ARI Contractor Report 96-76

# Technology Review of the Use of Continuous Speech Recognition for Language Training

C. Mazie Knerr
HumRRO

V. Melissa Holland
Army Research Institute

This report is published to meet legal and contractual requirements and may not meet ARI's scientific or professional standards for publication.

September 1996

United States Army Research Institute for the Behavioral and Social Sciences

Approved for pulbic release; distribution is unlimited.

# U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction
of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

## NOTICES

**DISTRIBUTION:** This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

**FINAL DISPOSITION:** This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

**NOTE:** The views, opinions and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

# REPORT DOCUMENTATION PAGE

| | | |
|---|---|---|
| 1. REPORT DATE (dd-mm-yy)<br>September 1996 | 2. REPORT TYPE<br>Contract Report | 3. DATES COVERED (from. . . to)<br>May-Oct 95 |

4. TITLE AND SUBTITLE

Technology Review of the Use of Continuous Speech Recognition for Language Training

5a. CONTRACT OR GRANT NUMBER
MDA903-93-D-0032 (D.O. 0031)

5b. PROGRAM ELEMENT NUMBER
0603007A

6. AUTHOR(S)

C. Mazie Knerr (HumRRO), V. Melissa Holland (ARI)

5c. PROJECT NUMBER
A793

5d. TASK NUMBER
2231

5e. WORK UNIT NUMBER
~~H~~C01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, Virginia 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Contractor Report 96-76

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This report is published to meet legal and contractual requirements and may not meet ARI's scientific or professional standards for publication.

14. ABSTRACT *(Maximum 200 words)*:

The Technology Review for Speech Recognition for Language Sustainment was an effort of the Special Operations Research. Development and Acquisition Center, the U.S. Army Research Institute and the Advanced Research Projects Agency in cooperation with the U.S. Army Special Operations Command Language Office. The review examined the state-of-the-art in continuous speech recognition (CSR) as it applies to foreign language training, sustainment, and enhancement for the Special Operations Forces (SOF).

The review addressed short-term, intermediate, and long-term goals for applying technology to SOF language training/sustainment needs. Presenters included developers of CSR systems with demonstrated interest in language education, ranging from industry to academia. In addition, participants discussed speech translation technology and its links to language training technologies. While the focus of the review was SOF, representatives of other military and government user groups also attended. It was held on August 2 and 3 1995, in Fayetteville, NC.

15. SUBJECT TERMS
Discrete speech recognition, Continuous speech recognition, Special operations forces, Speech translation technology, Language sustainment, Language training, Language education, Training technologies

| SECURITY CLASSIFICATION OF | | | | | |
|---|---|---|---|---|---|
| 16. REPORT<br>Unclassified | 17. ABSTRACT<br>Unclassified | 18. THIS PAGE<br>Unclassified | 19. LIMITATION OF ABSTRACT<br>Unclassified | 20. NUMBER OF PAGES<br>143 | 21. RESPONSIBLE PERSON (Name and Telephone Number)<br>Robert J. Seidel<br>(703) 617-8838 |

# Table of Contents

Page

**Summary** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

**Appendix A: Agenda** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A-1

## Appendix B: Speech Recognition Systems

Spoken Language Technology and Applications: State-of-the-Art
Clifford Weinstein (MIT Lincoln Laboratory) . . . . . . . . . . . . . . . . . . . . . . . . . . . . B-1

## Descriptions of Speech Recognition Systems

TriplePlay *Plus!*
Martin Rothenberg (Syracuse Language Systems, Inc.) . . . . . . . . . . . . . . . . . . . . . . . . B-35

Conversim™--A Dialog with a Native Speaker in a Multimedia Environment
William G. Harless (Interactive Drama, Inc.) . . . . . . . . . . . . . . . . . . . . . . . . . . . B-36

The Here and Now in Voice-Interactive Language Learning Systems
LTC Steve LaRocca and COL Woody Held (U.S. Military Academy) . . . . . . . . . . . . . B-37

Speech and Language Technology
Madeleine Bates and Sean Colbath (BBN Systems and Technologies) . . . . . . . . . . . . . B-38

Language Tutor and Bilingual Voyager System
Victor W. Zue, Joseph Polifroni, and Stephanie Seneff
(Massachusetts Institute of Technology) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B-39

Reprint of "Research and Development of Multilingual Conversational Systems" . . B-40

Voice Interactive Language Training System (VILTS)
Patti Price, Marikka Rypa, Leo Neumeyer, and George Chen (SRI);
Mike Valatka, Kathleen Egan, Helena Hughes, and Jacqueline Pogany (CIA) . . . . . . . B-46

Project LISTEN: A Reading Coach That Listens
Jack Mostow, Maxine Eskenazi, A. Hauptmann, B. Milnes,
and S. Roth (Carnegie Mellon University Robotics Institute) . . . . . . . . . . . . . . . . . . B-48

Entropic Speech Technology in Language Education
Jared Bernstein (Entropic Research Laboratory, Inc.) . . . . . . . . . . . . . . . . . . . . . . . B-49

Foreign Language Dialog System
John T. Lynch and Beth Carlson (MIT Lincoln Laboratory) . . . . . . . . . . . . . . . . . . . B-50

## Table of Contents (Cont.)

P

**Appendix C: Speech and Text Translation Systems**

Voice-to-Voice Machine Translation: Problems and Prospects
Susann Luperfoy (MITRE Corporation) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . (

**Descriptions of Speech Translation Systems**

JANUS: Spontaneous Speech to Speech Translation Environment Technology
Alex Waibel and Arthur E. McNair (Carnegie Mellon University) . . . . . . . . . . . . . . . C

Reprints of "Using Context in Machine Translation of Spoken Language" . . . . . . . C
and "JANUS: Towards Multilingual Spoken Language Translation" . . . . . . . . . . . C

Machine-Aided Voice Translation (MAVT) Advanced Development Model
Christine A. Montgomery, Bonnie Stalls, Robert E. Stumberger,
Naicong Li, Robert S. Belvin, Alfredo R. Arnaiz, and
Susan H. Litenatsky (Language Systems, Inc.) . . . . . . . . . . . . . . . . . . . . . . . . . . C

Reprint of "Machine-Aided Voice Translation (MAVT):
Advanced Development Model" . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C

Automatic English-to-Korean Text Translation of Naval Operational Reports
Young-Suk Lee, Dinesh Tummala, Stephanie Seneff, Cliff Weinstein,
and Jack Lynch (MIT Lincoln Laboratory) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C

Reprint of "CCLINC: System Architecture and Concept Demonstration
of Speech-to-Speech Translation for Limited-Domain Multilingual
Applications" . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C

Forward Area Language Converter (FALCON)
Daniel W. Smith, Jr. (XVIII Airborne Corps, Ft Bragg, NC) . . . . . . . . . . . . . . . . . . C

Briefing on "Forward Area Language Converter" . . . . . . . . . . . . . . . . . . . . . . . . . C

Multimedia Medical Language Translator
HMC(AW) Michael D. Hesslink and Captain Michael Valdez
(Naval Aerospace and Operational Medical Institute) . . . . . . . . . . . . . . . . . . . . . . . C

**Appendix D: References** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . I

**Appendix E: Revised List of Participants** . . . . . . . . . . . . . . . . . . . . . . . . . .

# TECHNOLOGY REVIEW:
# SPEECH RECOGNITION FOR LANGUAGE SUSTAINMENT

## Summary

The Technology Review for Speech Recognition for Language Sustainment was an effort of the Special Operations Research, Development and Acquisition Center (SORDAC), the U.S. Army Research Institute (ARI), and the Advanced Research Projects Agency (ARPA) in cooperation with the U.S. Army Special Operations Command (USASOC) Language Office. The purpose of the workshop was to review the state-of-the-art in continuous speech recognition as it applies to foreign language training, sustainment, and enhancement. Applications to Special Operations Forces (SOF) were the focus of presentations and discussions. The workshop was held on August 2 and 3, 1995, in Fayetteville, NC (Appendix A contains the agenda).

The review addressed short-term, intermediate, and long-term goals for applying technology to SOF language training/sustainment needs. It looked at what is available now or can be produced in the short term (1 year) with available technology; what can be done to meet SOF's needs in the mid-term by developing and exploiting advanced technologies (2 to 3 years out); and what to plan for from emerging technologies in longer-term research and development (5 to 20 years out). Presenters included major developers of continuous speech recognition systems with demonstrated interest in language education, ranging from industry to academia. They showed a variety of multilingual systems, some directly addressing language training and others readily adaptable to training and sustainment (Appendix B). In addition, participants discussed speech translation technology (Appendix C) and its links to language training technologies. While the focus of the review was SOF, representatives of other military and government user groups also attended (Appendix E lists the participants).

### First Day Focus: Training and Sustainment

The first major presentations of the day were by representatives of the Special Operations Forces (SOF) at Ft. Bragg. LTC Victor Kjoss, Chief of Training Division, DCSOPS, USASOC, overviewed the structure and missions of SOF and the role of foreign language skill in performing those missions. LTC H. Eugene Williams, 3rd Battalion, 1st Special Warfare Training Group, JFK Special Warfare Center and School, presented the school perspective on issues in initial language training. LTC Robert Brady, G-3 Special Forces Command, spoke on issues in language sustainment and enhancement from the perspective of the SOF Groups.

To begin the technology review, Dr. Cliff Weinstein of MIT Lincoln Laboratory overviewed applications of speech recognition technology (voice-based speaker identification, language identification, command and control, large vocabulary dictation, etc.) and described rapid growth over the past decade in the rates of recognition accuracy and the size of recognition vocabularies. For example, recognition of read speech, spoken continuously without pauses (known as continuous speech recognition) has progressed from vocabularies of

5K words to vocabularies of 60K words, with accuracy rates in the mid-90th percentile in highest performing recognizers.[1]

Nine system developers or groups then reviewed and demonstrated specific applications of speech recognition (Appendix B presents descriptions):

Dr. Martin Rothenberg, Syracuse Language Systems, Inc. (p. B-35)

Dr. William G. Harless, Interactive Drama, Inc. (p. B-36)

LTC Steve LaRocca and COL Woody Held, U.S. Military Academy (USMA), West Point (p. B-37)

Dr. Madeleine Bates and Mr. Sean Colbath, BBN Systems and Technologies (p. B-38)

Dr. Victor W. Zue, Dr. Joseph Polifroni, and Dr. Stephanie Seneff, Massachusetts Institute of Technology (MIT) (p. B-39)

Dr. Marikka Rypa, Dr. Patti Price, Dr. Leo Neumeyer, and Dr. George Chen, SRI; with Dr. Kathleen Egan, Ms. Helena Hughes, Dr. Mike Valatka, and Ms. Jacqueline Pogany, CIA Foreign Language Training Laboratory (p. B-46)

Dr. Jack Mostow and Dr. Maxine Eskenazi, Carnegie Mellon University (CMU) Robotics Institute (p. B-48)

Dr. Jared Bernstein, Entropic Research Laboratory, Inc. (p. B-49)

Dr. John T. Lynch and Dr. Beth Carlson, MIT Lincoln Laboratory (p. B-50)

The technologies applied ranged from lower-end systems using commercial off-the-shelf (COTS) recognizers that process discrete speech (single, fixed words and phrases) to higher-end systems using prototype recognizers that handle continuous speech (variable utterances, spoken naturally without pauses between words). The applications themselves varied from language tutoring to dictation to speech-activated database query.

The review included systems for purposes other than tutoring, as well as systems implemented in English rather than foreign languages, so as to demonstrate fully the pow speech recognition technology and to suggest the range of ways it might be deployed for foreign language sustainment. Languages in which recognizers were implemented includ English, Spanish, French, German, Italian, Japanese, Chinese, and Korean.

[1] Briefing charts and papers are presented in the appendices. References in parentheses cite the appendix ar where the material appears. Dr. Weinstein's briefing charts start on page B-1.

Discrete speech recognition engines have been available as COTS items for some time and can be purchased together with development kits that let system builders make their own speech-interactive applications. For example, the recognizer from Dragon Systems underlies two of the systems demonstrated: the commercial product TriplePlay *Plus!* from Syracuse Language Systems, which teaches core vocabulary in selected European languages, and the prototype instructional packages from Interactive Drama, which combine speech recognition with interactive video. The "talkie" language lessons designed by the USMA use the commercially available Aria Listener software to support vocabulary building as well as pronunciation training on foreign word pairs that are confusing to learners.

Continuous speech recognition (CSR) engines have been used largely in research prototypes. Several of the systems included in the review showed the power of CSR technology for authentic tasks in which users speak at natural rates, without pauses between words, with some freedom of expression, and without having to train the recognizer on their particular voice. Tasks included Wall Street Journal dictation (BBN), map navigation (MIT), and air travel information queries (MIT, BBN). For example, MIT's Voyager allows users to ask in Japanese the location of various sites within an American city. The system answers by highlighting the sites on a map of the city as well as by voicing a description of the location, in the user's choice of Japanese or English. Queries are unconstrained -- that is, users are not told in advance what to say or how to say it. Moreover, the system's estimation of what the user said is displayed on the screen. BBN's Air Travel Information System demonstrated a similar functionality for English questions about flight schedules and other travel information. The point was made that tasks like these can serve language sustainment by providing a simulated world in which the learner uses the target language to solve realistic problems typical of SOF missions.

The remaining CSR-based systems were developed specifically for language instruction, including the Voice Interactive Language Training System (VILTS) of SRI, the LISTEN tutor from Mostow at CMU, and the demonstrations by Bernstein from Entropic Research Laboratory as well as by Lynch and Carlson from Lincoln Laboratory. VILTS showed the precision of CSR technology for modeling learners' pronunciation and for diagnosing departures from native pronunciation in French. The system also showed how databases developed for speech recognition can be further exploited for listening comprehension, where learners can request to hear a particular word or idiom pronounced by different speakers in different utterance contexts. Mostow's LISTEN, developed to teach beginning readers of English, detects the words readers have trouble with and coaches them on the fly with hints and corrections as misreadings occur. Demonstrating the flexibility of the CSR approach, LISTEN generalizes to new texts without specific new training. SOF representatives viewing this demonstration suggested an immediate use for a foreign language LISTEN to coach personnel tasked with briefing foreign nationals in the native language. Bernstein demonstrated CSR programs for automatically assessing spoken language fluency as well as for communicative language instruction, in which learners describe a picture or direct an animated event in Spanish. Lincoln Laboratory demonstrated a lesson based on ARI's Military Language Tutor (MILT) in which the learner poses questions in Spanish to a modeled person who responds with prerecorded utterances in Spanish. The applications of both Lincoln Laboratory and Bernstein employ the HTK continuous speech recognizer marketed by Entropic, the highest performer in terms of accuracy rates in a sequence of ARPA competitions.

The discrete recognition systems of Syracuse Language Systems, Interactive Drama, and the USMA all run on conventional PC platforms (486 machines). They are intended as speaker independent (that is, individual users do not have to train the machine on their voices). The continuous recognizers, by contrast, run on workstations such as the Sparc, but some of these recognizers are being ported down. For example, the SPHINX continuous recognizer from CMU has been ported to a Pentium-based laptop running under Windows NT, as demonstrated by Mostow for the reading coach LISTEN. The HTK engine marketed by Entropic is being ported to a 486 PC running under Windows (scheduled for the end of 1995). This product includes a development kit that can be used to build new CSR applications. While designed as speaker independent, many of these recognizers perform better after a short period of adaptation to the individual speaker.

### Second Day Focus: Speech and Text Translation

Dr. Susann Luperfoy from MITRE overviewed the task of machine translation and what makes it hard. She analyzed the multiple aspects of language and communication that a computer program must consider in order to produce accurate translations (p. C-1).

Five system developers or groups then reviewed their translation systems. The systems were chosen to sample a range of approaches, from high-end, long-term solutions to low-end, short-term solutions. Two high-end systems addressed bidirectional, speech-to-speech translation of dialogues between speakers of different languages. These systems represent attempts to incorporate all the aspects of language and discourse described by Luperfoy: Waibel from CMU showed the JANUS system for translating between multiple language pairs, permitting any combination of English, German, or Spanish input (Korean and Japanese are under development), with English, German, Spanish, Korean, or Japanese output (p. C-27). Language Systems Inc. showed the machine-aided voice translator (MAVT), sponsored by Rome Laboratory and designed to translate between English and Spanish, with extensions underway to Arabic and Russian (p. C-48). Both systems incorporate an interlingual approach, in which the source language is translated into an abstract, universal semantic representation (an interlingua) before being converted to the target language. The interlingua provides maximum generalizability to new language pairs. In addition, both systems make the translation problem tractable by focusing on a single domain: meeting scheduling (Janus) and basic tactical interrogation (MAVT). Notably, Janus was designed to handle the disfluencies common in spontaneous speech (pauses, re-starts, and fillers like "um"). It collects large samples of real conversations around the target domain and then models the observed disfluencies so they can be systematically separated out when new conversations are processed. By training on large samples, Janus permits recognition and translation of new utterances that have not been specifically predicted.

Lincoln Laboratory demonstrated a bi-directional Korean-English translator, CCLINC, that works on text, thus eliminating the problem of speech recognition (p. C-56). This translator focuses on the domain of Naval operations messages and uses an interlingua for extendibility to new reports (p. C-57). These three high-end systems - Janus, MVAT, and CCLINC - currently run on workstations rather than PCs.

Two quick-term approaches to translation were also demonstrated. The FALCON (Forward Area Language Converter) uses a bilingual word list to perform word-for-word translation of a scanned-in foreign language document (p. C-63). Although the resulting English text is low on conventional measures of accuracy and readability, it usually gives enough information for the English-speaking soldier in the field to decide whether to forward the document to headquarters for full translation. The Army Materiel Command and the Army Research Laboratory are developing FALCON for the XVIII Airborne Corps. Currently available for French, it is being extended to other languages.

The Multimedia Medical Translator, demonstrated by HMC(AW) Hesslink, is a suite of nearly 2,000 prerecorded utterances in more than 40 languages, available on a CD-ROM disk for use in medical examinations (p. C-74). The user accesses the desired recordings by choosing from menus of English questions and expressions. The corresponding foreign language utterances are then played by the device. Questions are designed to elicit yes-no answers or pointing responses. Developed by the Naval Aerospace and Operational Medical Institute, this program is being used by Naval health care staff supporting U.N. operations in the former Yugoslavia. The program was recently extended to training in mine clearing operations. Both the Multimedia Medical Translator and FALCON run on a PC, laptop, or notebook equivalent.

Systems for translation were included in the review, first, because SOCOM has a documented requirement for translation, both text- and speech-based; second, because many of the components developed for translation can also support language training and sustainment. Cooperative agreements to share technologies already exist between ARI and the various agencies that support translation work.

## Conclusions

Government participants in the review included scientists as well as end users representing SOF, the Army Research Institute, ARPA, the Army Intelligence Center and School, the Defense Language Institute, the Deputy Chief of Staff for Intelligence (HQDA), Deputy Chief of Staff for Operations (HQDA), the Army Research Laboratory, Army Training and Doctrine Command, Army Research Office, CIA, NSA, DCI Foreign Language Committee, and Rome Laboratory (Air Force), among other agencies (Appendix E). Government representatives generally agreed that the core technologies demonstrated at the review - discrete and continuous speech recognition - were sufficiently mature to support a robust language sustainment tutor with which learners can interact by speaking. Moreover, it was agreed that these technologies appear suitable for both pronunciation training and practice of conversational, communicative tasks in target languages. Both commercial and research demonstrations were credible in that most permitted new and unpracticed users to interact with the system without significant performance deficits.

At the same time, it was agreed that applied research and development are needed to shape the core technologies into a product useful to SOF. Commercially available software, while useful for global language training, does not address SOF-specific tasks and vocabulary, nor is it available in the more difficult languages critical to SOF (e.g., Arabic, Korean, Thai). Moreover, commercial language learning products currently use discrete recognition algorithms and do not exploit the power of CSR to process spontaneous, variable utterances. Similarly, research prototypes, many of which do employ CSR to train language learning skills, are not available in high-priority languages, nor do they address task domains of concern to SOF. Plans were made, then, to develop a short-term (1-year) language sustainment tutor using discrete speech recognition and a medium-term (2-year) tutor using continuous speech recognition, both addressing SOF-critical languages and tasks. Beginning in FY96, this development is to be supported by a joint program involving SOCOM, ARPA, and ARI, working through the SOF Language Office and guided by specific input from the SOF Groups, NAVSOC, and AFSOC.

UNITED STATES SPECIAL OPERATIONS COMMAND

Technology Review:
Special Operations Forces (SOF)
Speech Recognition for Language Sustainment

# Appendix A:
# Agenda

AIRBORNE

ARMY RESEARCH INSTITUTE

## TECHNOLOGY REVIEW: SPECIAL OPERATIONS FORCES (SOF) SPEECH RECOGNITION FOR LANGUAGE SUSTAINMENT

### AGENDA

Wednesday - 2 August 1995

0730 Registration Opens - Continental Breakfast

0830 Introduction - Melissa Holland (ARI)

Gil Buhrmann (Office of Special Technology)
Allen Sears (ARPA Human Language Systems and Human Computer Interactions)
Mike Sanders (ARI, Ft. Bragg)

0850 SOF Language Training and Sustainment

Overview
LTC Kjoss, SOF Language Office (Interservice)
School Perspective: Initial Language Training
LTC Williams (JFK Special Warfare Center and School)
Groups Perspective: Language Sustainment
LTC Brady (US Army SF Command)
Questions for SOF

1000 Break

1015 Speech Recognition (SR) State-of-the-Art.
Cliff Weinstein (Lincoln Lab)

1045 Introduction to the Systems: SR for Language Training/Sustainment - Set 1 and Set 2 Systems

1230 Lunch

## **AGENDA** (Cont.)

Wednesday - 2 August 1995 (Cont.)

| | | |
|---|---|---|
| 1330 | Demonstrations of Set 1 Systems | |
| 1510 | Break | |
| 1525 | Demonstrations of Set 2 Systems | |
| 1710 | Summary and Announcements | - Melissa Holland (ARI)<br>- Mazie Knerr (HumRRO) |
| 1730 | Reception with Cash Bar | |
| 1900 | Dinner | |

# AGENDA (Cont.)

## Thursday - 3 August 1995

0730 Continental Breakfast (General Meeting Room)

0830 Introduction - Melissa Holland (ARI)

0835 Speech Translation: Problems and Prospects - Susann Luperfoy (MITRE)

0900 Introduction to the Systems: Translation and Speech Recognition - Set 3 Systems

0945 Break

1000 Demonstrations: Speech Translation Systems - Set 3 Systems

1140 Discussion and Summary - Robert J. Seidel (ARI)

1200 Adjourn general meeting

**Demos from August 3 are available until 1245**

*Notes: Meetings on August 3*

- *ARPA developers meet with Allen Sears from 0700 - 0830 (Palais Room)*
- *Government meeting with SOF representatives from 1330 - 1530 (General Meeting Room)*

UNITED STATES SPECIAL OPERATIONS COMMAND

Technology Review:
Special Operations Forces (SOF)
Speech Recognition for Language Sustainment

# Appendix B: Speech Recognition Systems

AIRBORNE

ARMY RESEARCH INSTITUTE

UNITED STATES SPECIAL OPERATIONS COMMAND

# Technology Review:
# Special Operations Forces (SOF)
# Speech Recognition for Language Sustainment

## Dr. Clifford Weinstein's Presentation
## "Spoken Language Technology and Applications: State-of-the-Art"

AIRBORNE

ARMY RESEARCH INSTITUTE

# SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS: STATE-OF-THE-ART

PRESENTATION AT THE
SPECIAL OPERATIONS FORCES TECHNOLOGY REVIEW ON
SPEECH RECOGNITION FOR LANGUAGE SUSTAINMENT
FAYETTEVILLE, NC
2 AUGUST 1995

CLIFFORD WEINSTEIN
MIT LINCOLN LABORATORY
244 WOOD STREET
LEXINGTON, MA 02173-9108
PHONE: 617-981-7491; FAX 617-981-0186
E-MAIL: CJW@SST.LL.MIT.EDU

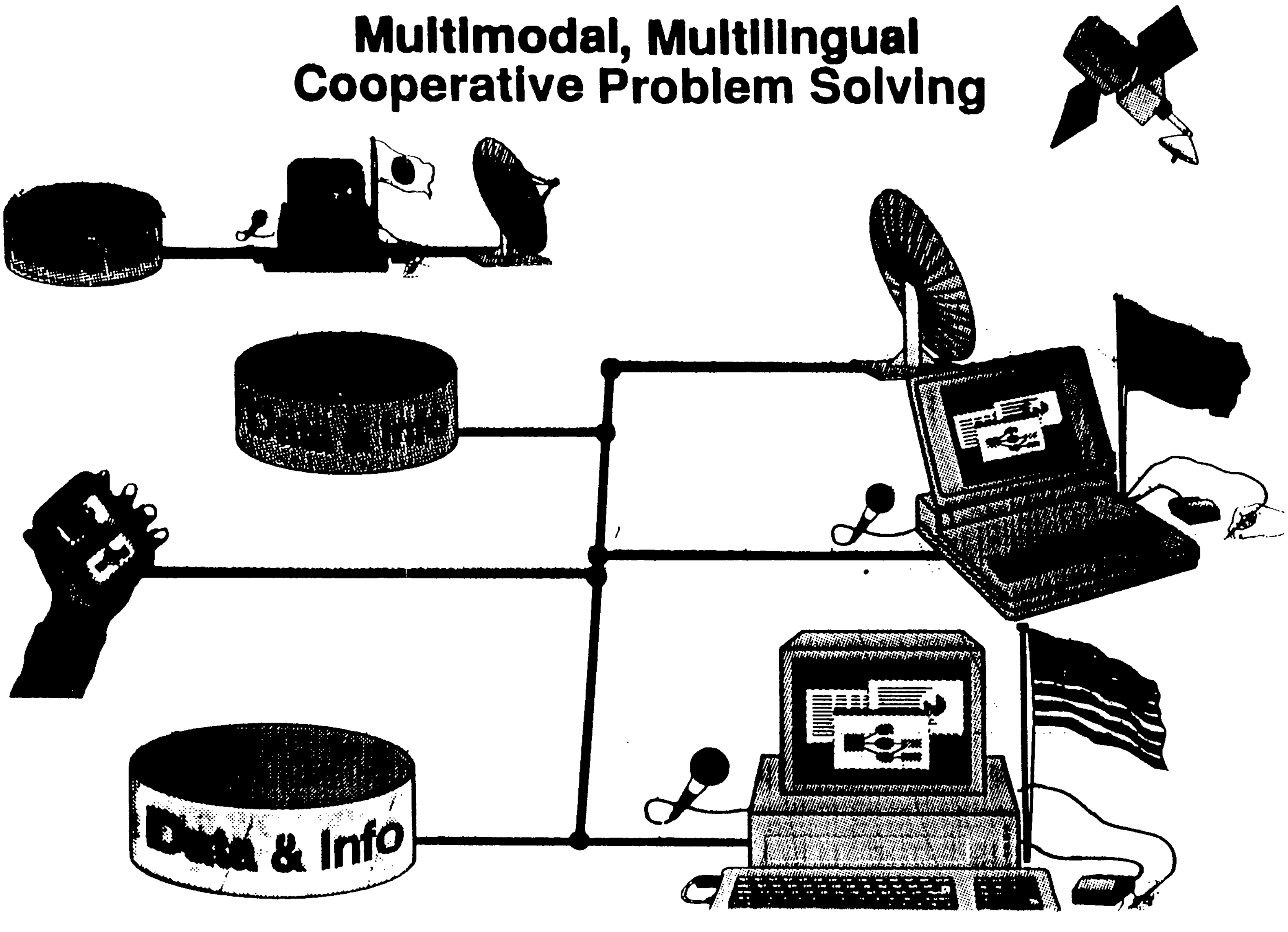
# BACKGROUND: WORKSHOPS AND STUDIES ON SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS

- MILITARY SPEECH TECH 87: SPEECH IN AFTI F16, ARMY HELICOPTERS
- ARPA ISAT 88, SPOKEN LANGUAGE SYSTEMS
  - FOCUS ON ADVANCES IN TECHNOLOGY VIA HMM AND BEYOND
- NATO RSG10 STUDY, COMPLETED 91
  - "OPPORTUNITIES FOR ADVANCED SPEECH PROCESSING IN MILITARY COMPUTER-BASED SYSTEMS"
- ARPA ISAT 92, MULTIMODAL LANGUAGE-BASED SYSTEMS
  - DIALOG SYSTEMS VISION AND APPLICATIONS
- NAS 93, VOICE COMMUNICATIONS BETWEEN HUMANS AND MACHINES
  - COLLOQUIUM AND BOOK, TECHNOLOGY AND APPLICATIONS
- ARPA SPOKEN LANGUAGE TECHNOLOGY & APPLICATIONS DAY, SLTA '93
  - DEMOS OF RECOGNITION AND UNDERSTANDING; USERS' PANEL
- ARMY ARO/TRADOC 94 WORKSHOP ON NATURAL LANGUAGE AND SPEECH RECOGNITION TECHNOLOGY
  - TECHNOLOGY, APPLICATIONS, USERS; C2, TRAINING, AND TRANSLATION
- ARMY ARO/TRADOC MAY 95 SPOKEN HUMAN-MACHINE DIALOGUE WORKSHOP
  - FOCUS ON APPLICATIONS TO TRAINING, ASSISTANCE AND OPERATIONS
  - ADDRESS BATTLE LAB MISSIONS, E.G. BATTLE COMMAND, COMBATSERVICE SUPPORT
- SOF WORKSHOP: SPEECH RECOGNITION FOR LANGUAGE SUSTAINMENT, AUG. 2-3, 1995
  - REQUIREMENTS AND TECHNOLOGY REVIEW, DEMOS
  - FOCUS ON APPLICATIONS TO LANGUAGE RAINING, SUSTAINMENT, TRANSLATION
- ARPA SOFTWARE TECHNOLOGY AND INFORMATION SYSTEMS SYMPOSIUM (STISS95), AUGUST 28-31, 1995

## SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS: STATE-OF-THE-ART

➡
- VISION AND CHALLENGE
- TECHNOLOGY ISSUES AND PROGRESS
- MILITARY AND GOVERNMENT APPLICATIONS
- CURRENT THRUSTS IN DIALOGUE SYSTEMS, LANGUAGE EDUCATION, AND SPEECH TRANSLATION
- PLANS AND THE FUTURE

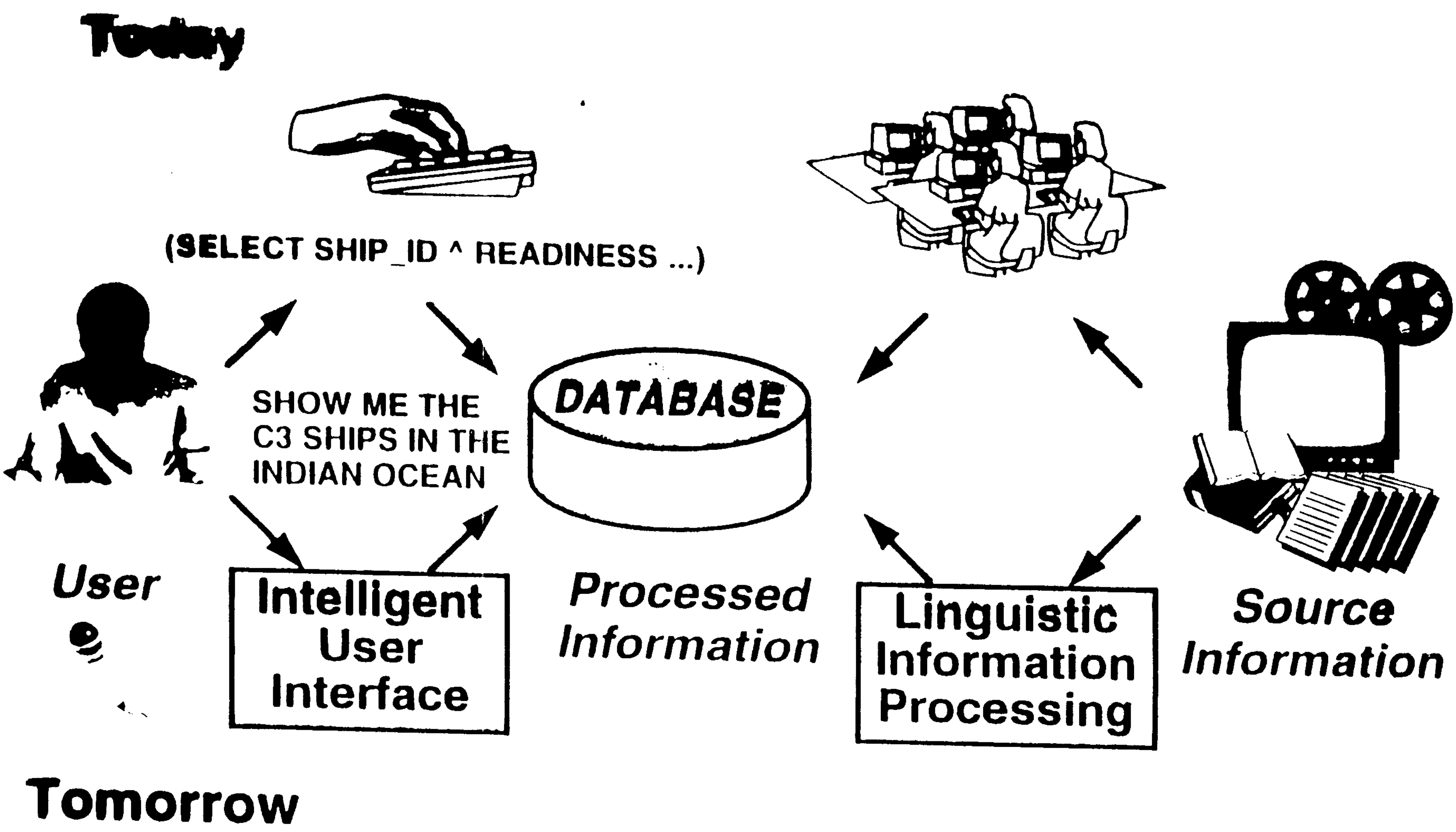
# Multimodal, Multilingual Cooperative Problem Solving



# Example Application: Mission Planning and Coordination

- Crisis in former Yugoslavia requires the deployment of UN peace-keeping forces
- Mission planning staff from many nations work cooperatively and collaboratively
- Decisions are time-critical, requiring instantaneous access to vast amounts of information
- All databases are linked; information is conveyed in appropriate ways, translated if necessary
- System permits convenient access and manipulation of information, using various input/output capabilities

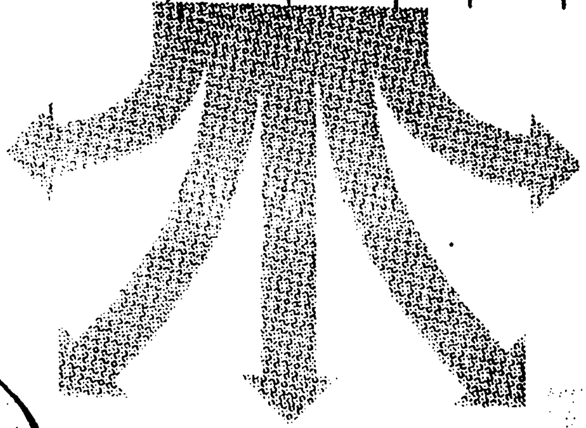
# Providing Information for the User



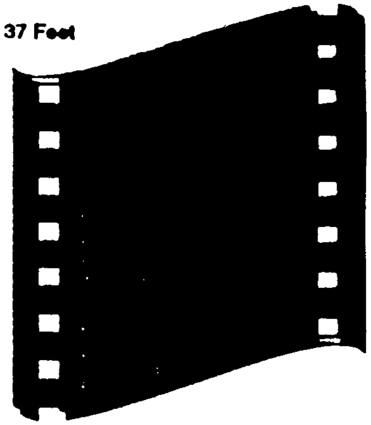
Today
(SELECT SHIP_ID ^ READINESS ...)
SHOW ME THE C3 SHIPS IN THE INDIAN OCEAN
User
Intelligent User Interface
DATABASE
Processed Information
Linguistic Information Processing
Source Information
Tomorrow

# Example of Intelligent Multimodal and Multimedia Mission Planning

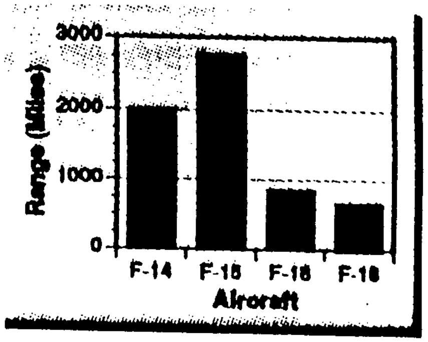
| ID | Name | Speed | Range | Thrust | Crew | Length | Wingspan |
|---|---|---|---|---|---|---|---|
| F-14 | Tomcat | Mach 2.34 | 2000 Miles | 14,000x2 | 2 | 62 Feet | 64 Feet |
| F-15 | Eagle | Mach 2.5 | 2765 Miles | 29,000x2 | 2 | 63 Feet | 42 Feet |
| F-16 | Falcon | Mach 2.0+ | 850 Miles | 28,900x1 | 1 | 49 Feet | 31 Feet |
| F-18 | Hornet | Mach 1.8 | 680 Miles | 16,000x2 | 1 | 56 Feet | 37 Feet |



| Aircraft | Range |
|---|---|
| F-14, Tomcat | 2000 Miles |
| F-15 Eagle | 2765 Miles |
| F-16 Falcon | 850 Miles |
| F-18 Hornet | 680 Miles |



The mission parameters exceed the capabilities of an F-18 aircraft. Its range is limited to six-hundred eighty miles.



# Grand Challenge

- **Develop language-based technologies for:**
  - Rapid, effortless human-machine (and human-human) communication
  - Machine processing of vast quantities of spoken and written material
- Demonstrate such systems in application areas relevant to the DoD

# Key Ideas

- Interactivity
- Multimodal input, multimedia output
- Multiple languages
- Automated information extraction and summarization

# SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS: STATE-OF-THE-ART

- VISION AND CHALLENGE
- TECHNOLOGY ISSUES AND PROGRESS
- MILITARY AND GOVERNMENT APPLICATIONS
- CURRENT THRUSTS IN DIALOGUE SYSTEMS, LANGUAGE EDUCATION, AND SPEECH TRANSLATION
- PLANS AND THE FUTURE

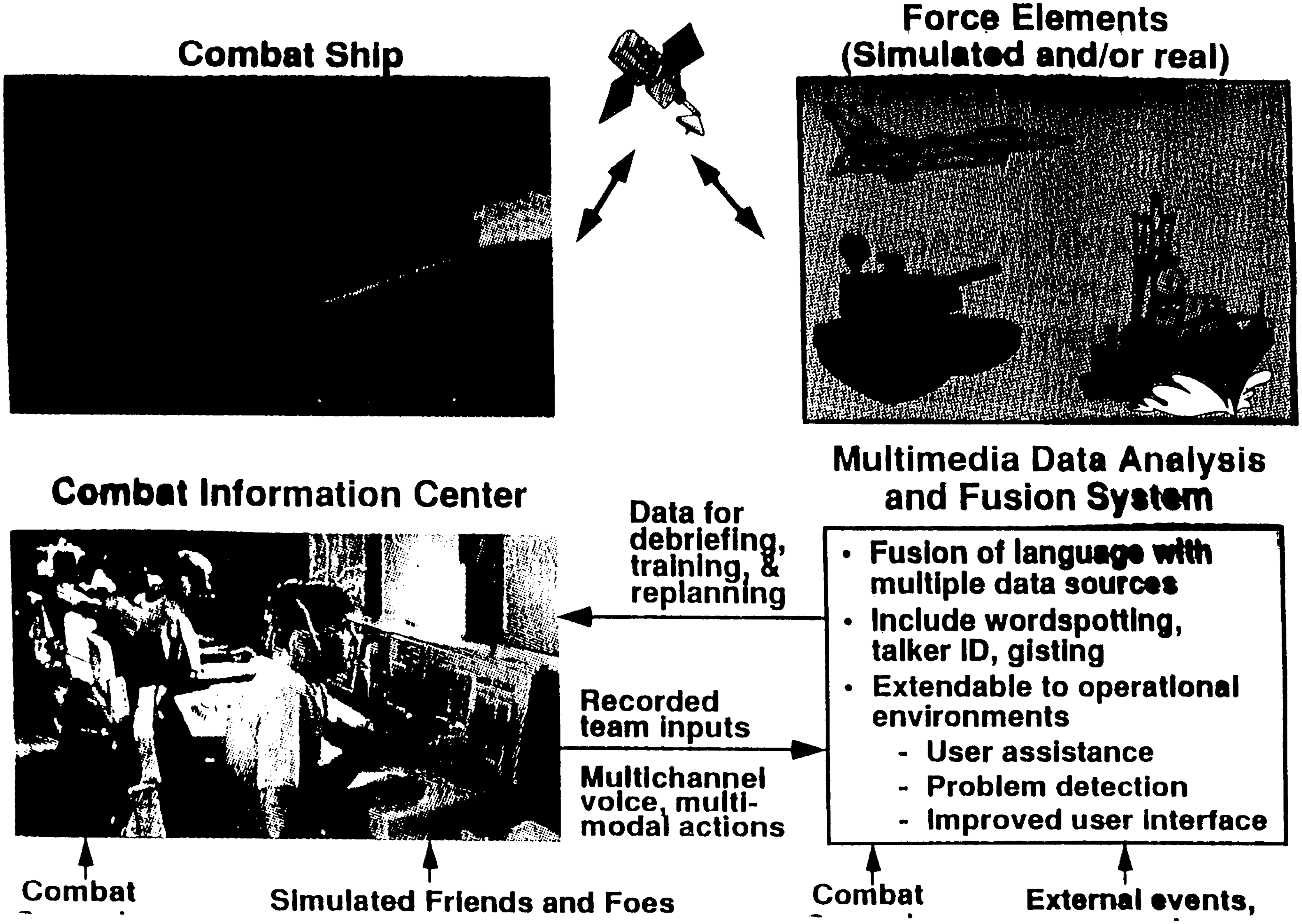
# The Information Analyst's Computer

- Users: Information analyst, mission planner, reporter, researcher, student, librarian
- Functions
  - Selective information retrieval and extraction from multilingual and multimedia sources
  - Text and speech generation in user's language
  - Summarization, abstraction, highlighting
  - Translation, rough and fine (human-aided)
  - Voice transcription (e.g., broadcasts, interviews)
  - Document creation (dictation and drawing)

# The Agent's Computer

- Users:  FBI agent, special forces agent, police
- Functions:
  - Voice check-in, verification
  - Data or report entry
  - Data access (license plate check, description-based check)
  - Map and direction information
  - Covert communication
  - Simple translation

# Combat Team Tactical Training



Combat Ship

Force Elements (Simulated and/or real)

Combat Information Center

Multimedia Data Analysis and Fusion System

- Fusion of language with multiple data sources
- Include wordspotting, talker ID, gisting
- Extendable to operational environments
  - User assistance
  - Problem detection
  - Improved user interface

Data for debriefing, training, & replanning

Recorded team inputs

Multichannel voice, multi-modal actions

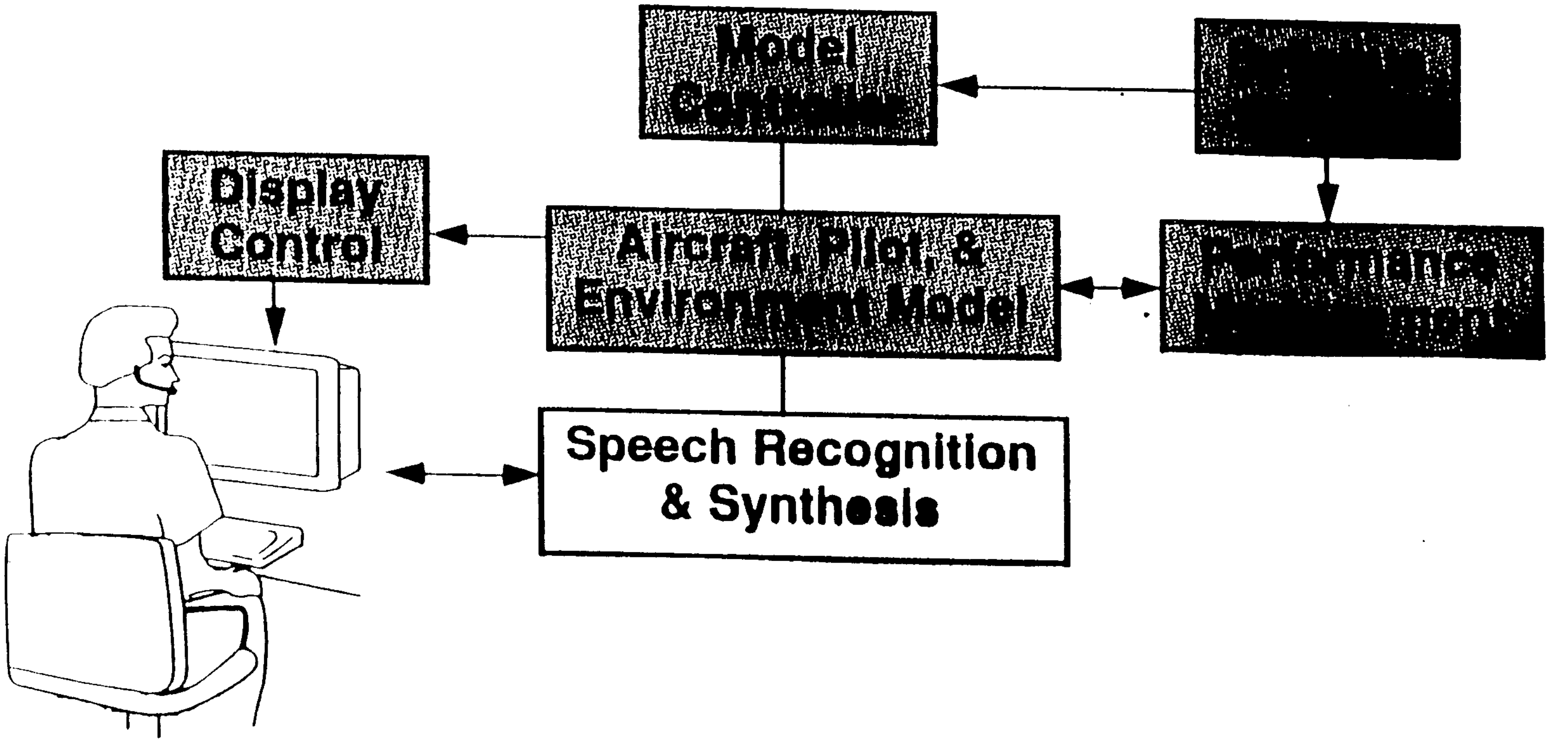
Combat

Simulated Friends and Foes

Combat

External events,

# Air Traffic Control Training and Automation



Model Controller

[illegible]

Display Control

Aircraft, Pilot & Environment Model

Performance Assessment

Speech Recognition & Synthesis

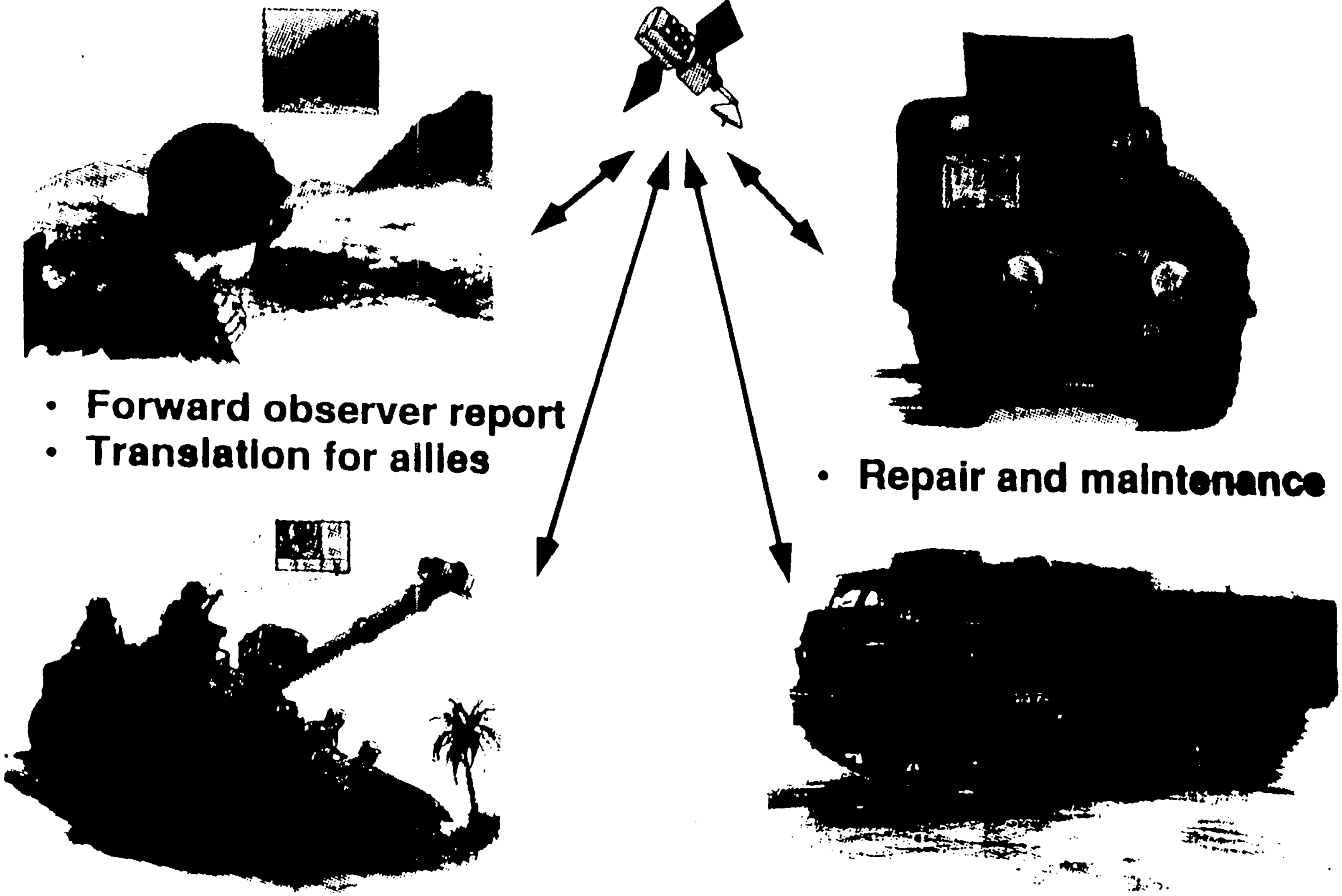
## Training

- Emulation of pseudo-pilots
- Scenario driven simulation
- Automated session analysis (including speech recognition)

## Automation

- Active control (e.g., flight ID recognition)
- Processing of multi-channel voice and sensor data to evaluate automation aids
- Gisting for conflict detection

# Command-and-Control on the Move (C2OTM)



- Forward observer report
- Translation for allies
- Repair and maintenance
- Information access/display

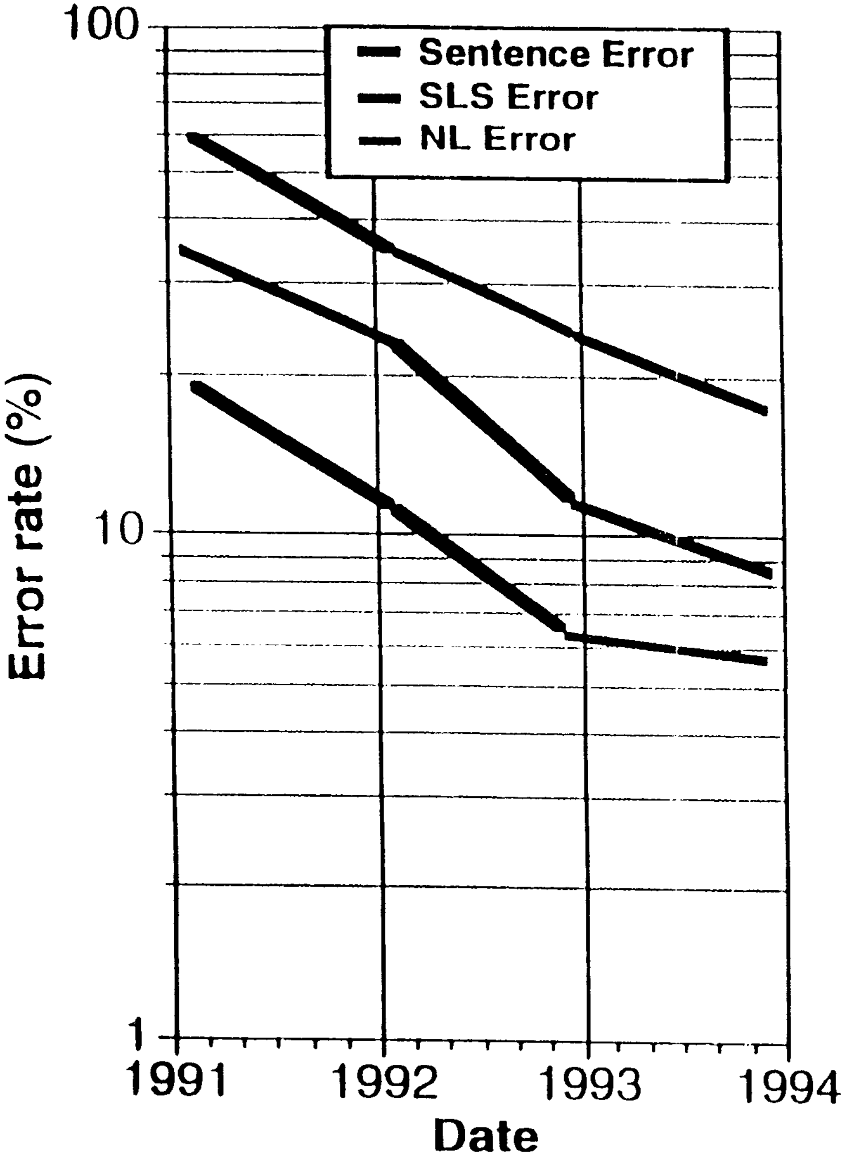
# SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS: STATE-OF-THE-ART

- VISION AND CHALLENGE
- TECHNOLOGY ISSUES AND PROGRESS
- ➡ MILITARY AND GOVERNMENT APPLICATIONS
- CURRENT THRUSTS IN DIALOGUE SYSTEMS, LANGUAGE EDUCATION, AND SPEECH TRANSLATION
- PLANS AND THE FUTURE

# SPOKEN LANGUAGE DEMONSTRATION SYSTEMS*

- AIR TRAVEL INFORMATION SYSTEM (MULTILINGUAL)
- URBAN DIRECTION ASSISTANCE (MULTILINGUAL)
- MULTILINGUAL INTEGRATED INFORMATION ASSISTANCE
- SPEECH-TO-SPEECH TRANSLATION
  - ENGLISH, SPANISH, GERMAN, FRENCH, SWEDISH, JAPANESE, KOREAN, ARABIC
  - LIMITED DOMAINS: MEETING PLANNING, INTERROGATION
- VOICE-CONTROLLED FLIGHT SIMULATOR
- FLIGHT ID RECOGNITION FOR AIR TRAFFIC CONTROL
- VOICE BANKING
- FOREIGN LANGUAGE EDUCATION
- READING COACH
- C4I INFORMATION INPUT AND ACCESS
- LOGISTICS PLANNING
- OFFICE MANAGER
- MULTILINGUAL DICTATION
- 40,000-WORD CONTINUOUS SPEECH RECOGNITION

# State of the Art: Example in the ATIS Domain

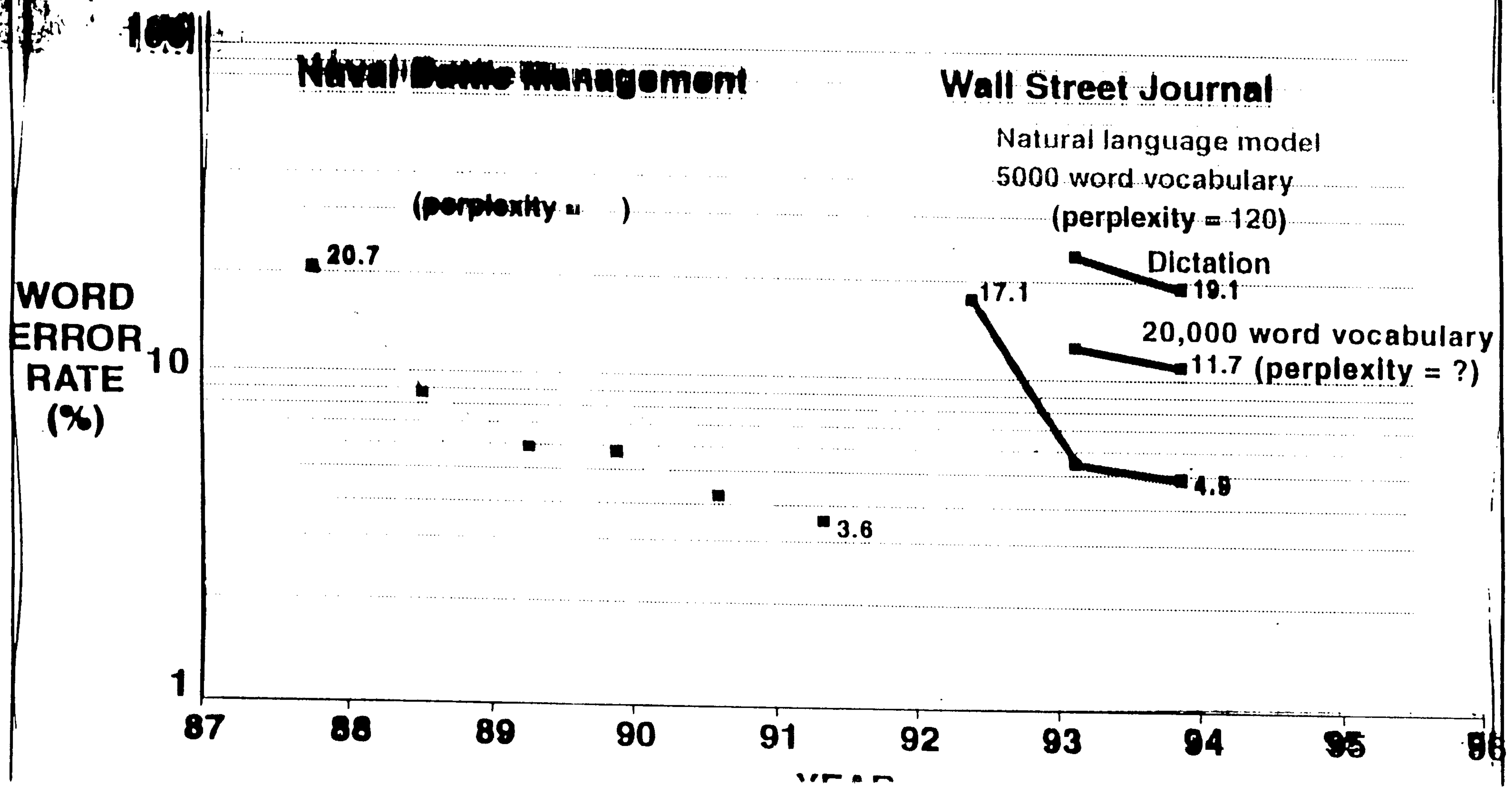


- Error rate cut in half every two years
- Many more sentences understood than recognized completely - complete NL analyses not required
- Speech understanding performance supports effective human-machine dialogue

# Progress In Speech Recognition

ARPA

*Software and Intelligent Systems Technology*



# DIMENSIONS OF DIFFICULTY
# FOR SPOKEN LANGUAGE APPLICATIONS

- TASK-RELATED
  - SPEAKER INDEPENDENCE
  - VOCABULARY SIZE (AND CONFUSABILITY)
  - SPEAKING MODE (ISOLATED, CONTINUOUS, WORDSPOTTING)
  - AVAILABILITY OF TRAINING DATA IN THE TASK DOMAIN
  - GRAMMAR PERPLEXITY
  - GRAMMAR COMPLEXITY AND AMBIGUITY
  - KNOWLEDGE AND PREDICTABILITY OF USER GRAMMAR AND VOCABULARY
- USER-RELATED
  - USER TOLERANCE OF ERRORS AND AVAILABILITY OF ERROR RECOVERY MECHANISMS
  - USER COOPERATION AND TRAINING
  - USER PRONUNCIATION AND ACCENT
  - USER STRESS
- ENIVIRONMENT-RELATED
  - ACOUSTIC BACKGROUND
  - CHANNEL AND MICROPHONE QUALITY

# Goals

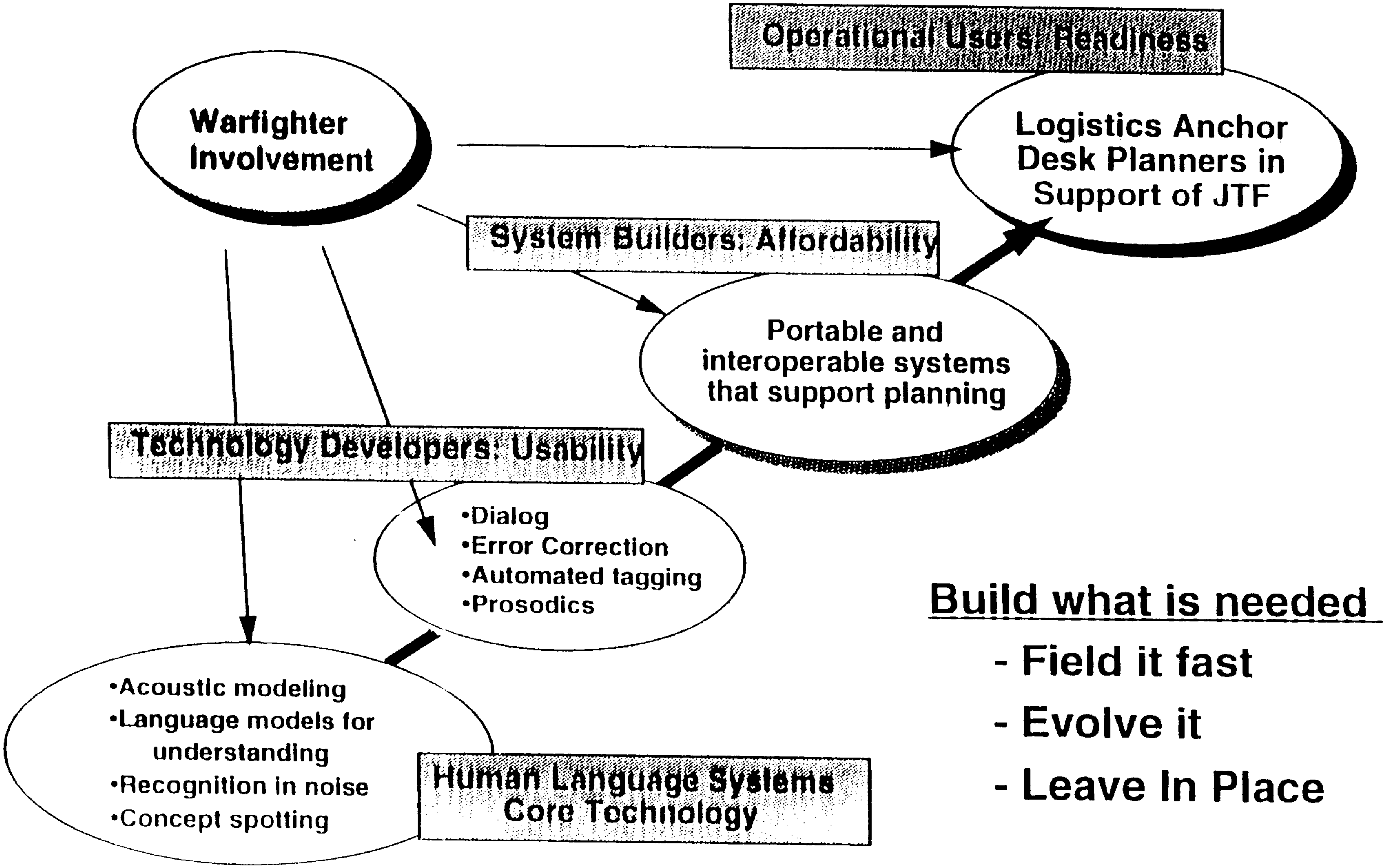
Within ten years, multimodal language-based technology can enable users to:

- Solve problems interactively with computers in constrained domains
- Prepare reports and documents in open domains
- Retrieve and analyze vast quantities of information
- Translate from one language to another
  - Text: Automatic, browsing quality
  - Speech: Interactive

This enabling technology will become an integral part of many DoD applications, including:

- Mission planning
- Command & control (on the move)
- Simulation and training
- Maintenance

# HLS Technology Transfer Strategy


Operational Users: Readiness
Warfighter Involvement
Logistics Anchor Desk Planners in Support of JTF
System Builders: Affordability
Portable and interoperable systems that support planning
Technology Developers: Usability
•Dialog
•Error Correction
•Automated tagging
•Prosodics
•Acoustic modeling
•Language models for understanding
•Recognition in noise
•Concept spotting
Human Language Systems Core Technology

<u>Build what is needed</u>

- Field it fast
- Evolve it
- Leave In Place

12/1/93

# TECHNOLOGY TRANSFER

- THE OPPORTUNITY:
  - GROWING INFORMATION-DEPENDENCE IN MILITARY OPERATIONS AND TRAINING CONTINUES TO INCREASE THE NEED FOR LANGUAGE-BASED HUMAN/COMPUTER INTERACTION
  - VERY BROAD RANGE OF USEFUL APPLICATIONS
  - VERY BROAD RANGE OF DIFFICULTY/TIME/COST TO ACHIEVE SUCCESSFUL APPLICATIONS
- KEY ISSUES
  - USERS NEED PORTABLE PRODUCTS, COTS IF POSSIBLE, WHICH THEY CAN ADAPT TO THEIR APPLICATIONS
  - WE MUST NARROW THE GAP BETWEEN THE USERS AND THE STATE-OF-THE-ART
- APPROACHES
  - DEVELOP EASILY PORTABLE PRODUCTS (SOFTWARE & HARDWARE)
  - SIMULATE USER ENVIRONMENT AND SHOW COMPELLING DEMONSTRATIONS
  - INTEGRATE, TEST, ITERATE, THEN [illegible]

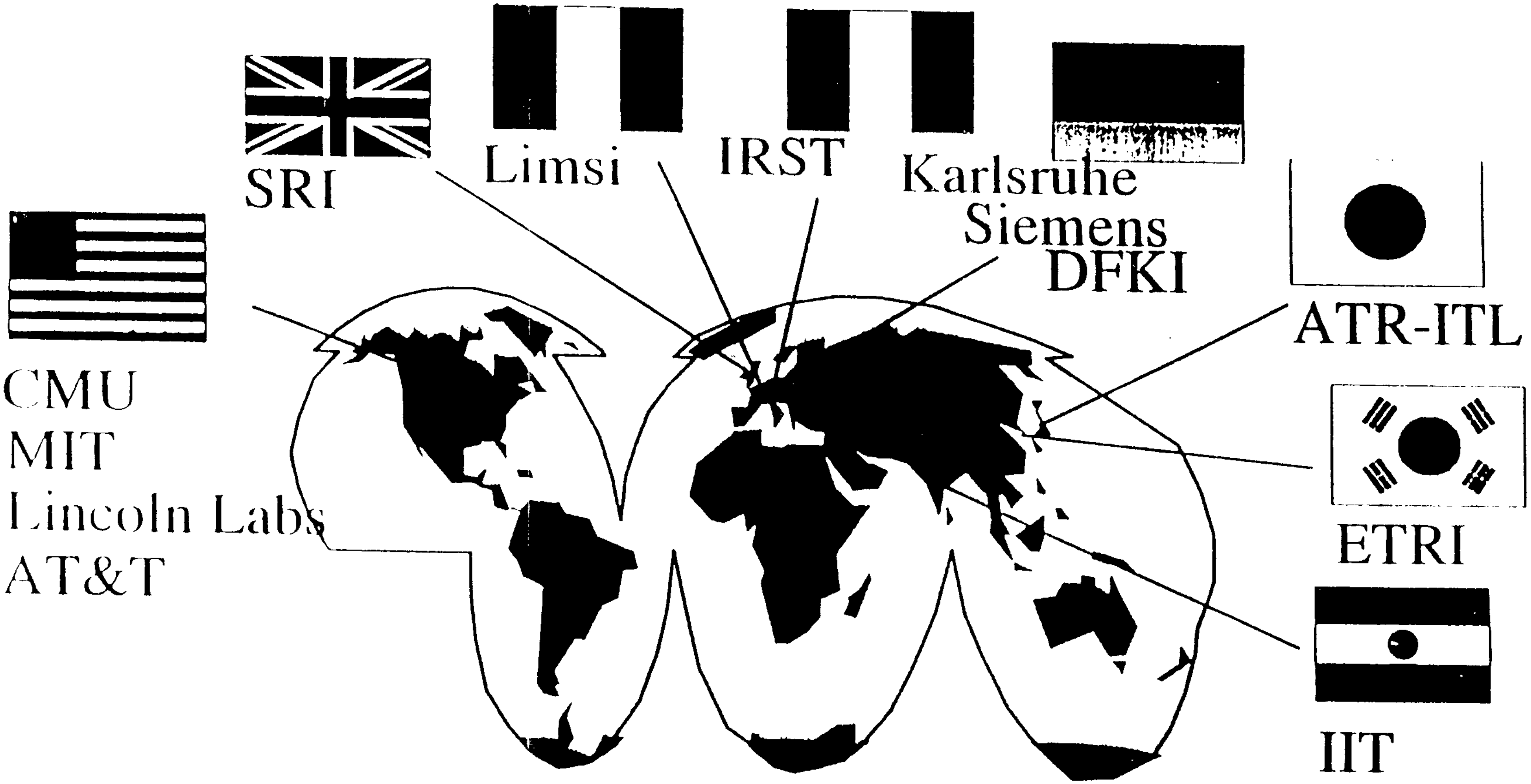
# HUMAN-MACHINE INTERACTIONS IN 2020

- VISION AND CHALLENGE
- MILITARY AND GOVERNMENT APPLICATIONS
- CURRENT TECHNOLOGY AND APPLICATIONS THRUSTS
- ➡ PLANS AND THE FUTURE

# CSTAR-II: CONSORTIUM FOR SPEECH TRANSLATION ADVANCED RESEARCH

- INFORMAL CONSORTIUM, NOT A FUNDING ORGANIZATION
- MET JUNE 1994 IN MUNICH; APRIL 1995 IN PITTSBURGH
- ATTENDEES: SIEMENS (GERMANY), ATR (JAPAN), CMU (U.S.), ETRI (KOREA), SRI/UK, IRST (ITALY), LIMSI (FRANCE), SAARBRUCKEN (GERMANY, MIT/LCS (U.S.), MIT/LL (U.S.), ATT-BL (U.S.)
- LANGUAGES: ENGLISH, GERMAN, KOREAN, JAPANESE, SPANISH, ITALIAN, FRENCH, PINYIN, SWEDISH
- EACH PARTICIPANT AGREES TO BUILD A HALF-DUPLEX SYSTEM
  - E.G., ENGLISH SPEECH TO GERMAN TEXT; KOREAN SPEECH TO JAPANESE TEXT
  - PLANNED ETRI PARTICIPATION: KOREAN TO ENGLISH, KOREAN TO JAPANESE
- OTHER C-STAR MEMBERS ARE AFFILIATES (CONDUCT RESEARCH, PRESENT AT WORKSHOPS, MAY BUILD SYSTEM COMPONENTS)
- INITIAL TASK IS MEETING SCHEDULING (1,000 - 2,000 WORDS)
- NEXT WORKSHOP IN JAPAN, SEPT. 1996 (APPROX.)

# C-STAR II



CMU
MIT
Lincoln Labs
AT&T

SRI

Limsi

IRST

Karlsruhe
Siemens
DFKI

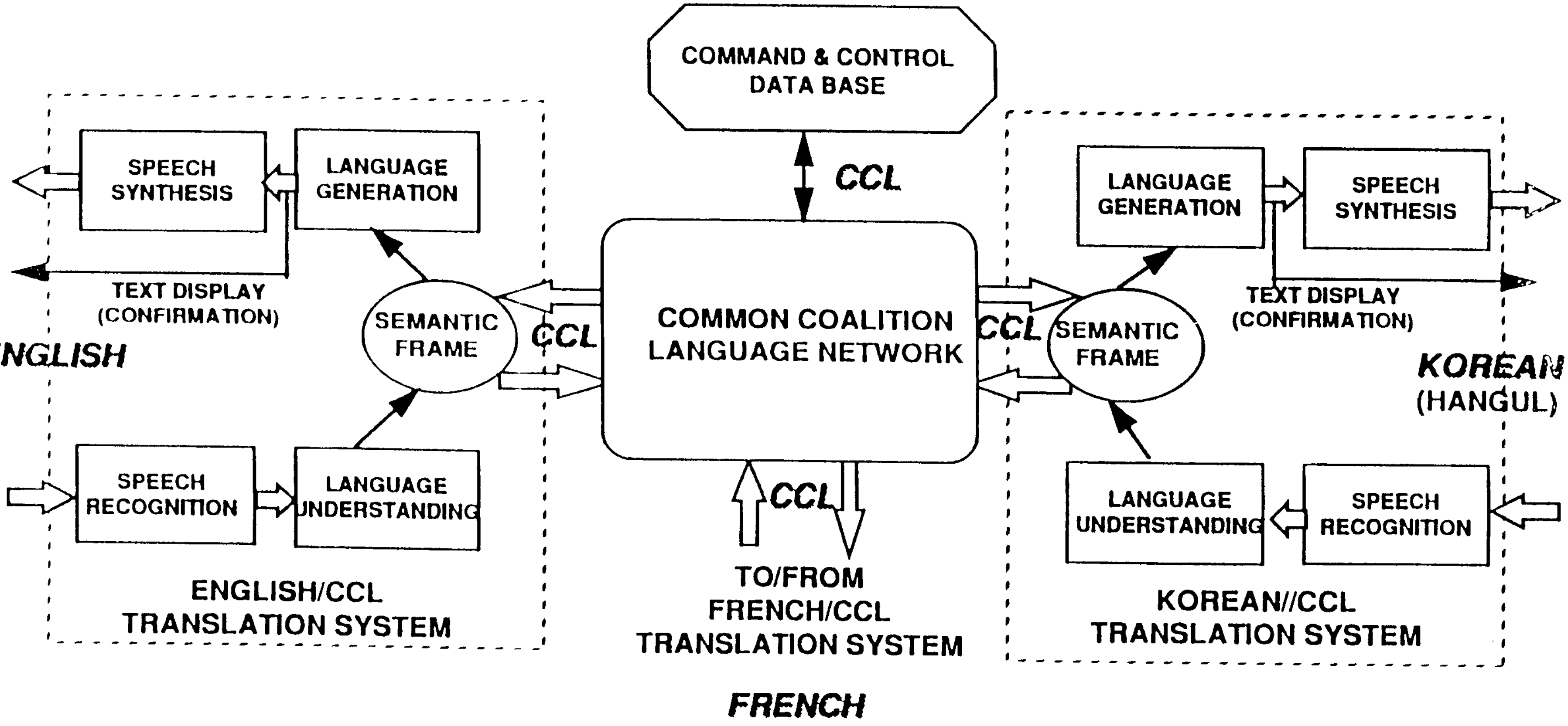
ATR-ITL

ETRI

IIT

*C-STAR II*

# SYSTEM STRUCTURE FOR MULTILINGUAL SPEECH-TO-SPEECH TRANSLATION



COMMAND & CONTROL DATA BASE

CCL

COMMON COALITION LANGUAGE NETWORK

SPEECH SYNTHESIS

LANGUAGE GENERATION

TEXT DISPLAY (CONFIRMATION)

SEMANTIC FRAME

CCL

SPEECH RECOGNITION

LANGUAGE UNDERSTANDING

ENGLISH

ENGLISH/CCL TRANSLATION SYSTEM

CCL

TO/FROM FRENCH/CCL TRANSLATION SYSTEM

FRENCH

LANGUAGE GENERATION

SPEECH SYNTHESIS

TEXT DISPLAY (CONFIRMATION)

CCL

SEMANTIC FRAME

KOREAN (HANGUL)

LANGUAGE UNDERSTANDING

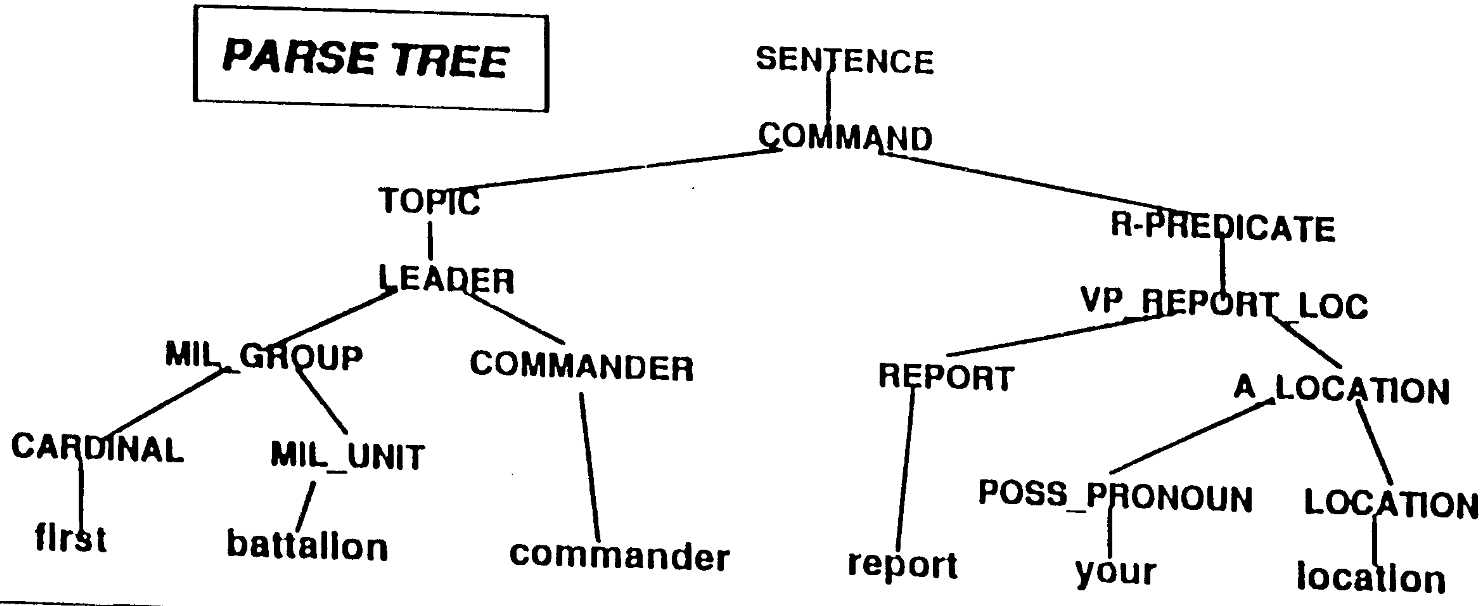
SPEECH RECOGNITION

KOREAN/CCL TRANSLATION SYSTEM

# EXAMPLE OF SENTENCE INTERPRETATION

- SENTENCE INPUT: "FIRST BATTALION COMMANDER REPORT YOUR LOCATION"

**_PARSE TREE_**



SENTENCE

COMMAND

TOPIC

LEADER

MIL_GROUP

CARDINAL

first

MIL_UNIT

battalion

COMMANDER

commander

R-PREDICATE

VP_REPORT_LOC

REPORT

report

A_LOCATION

POSS_PRONOUN

your

LOCATION

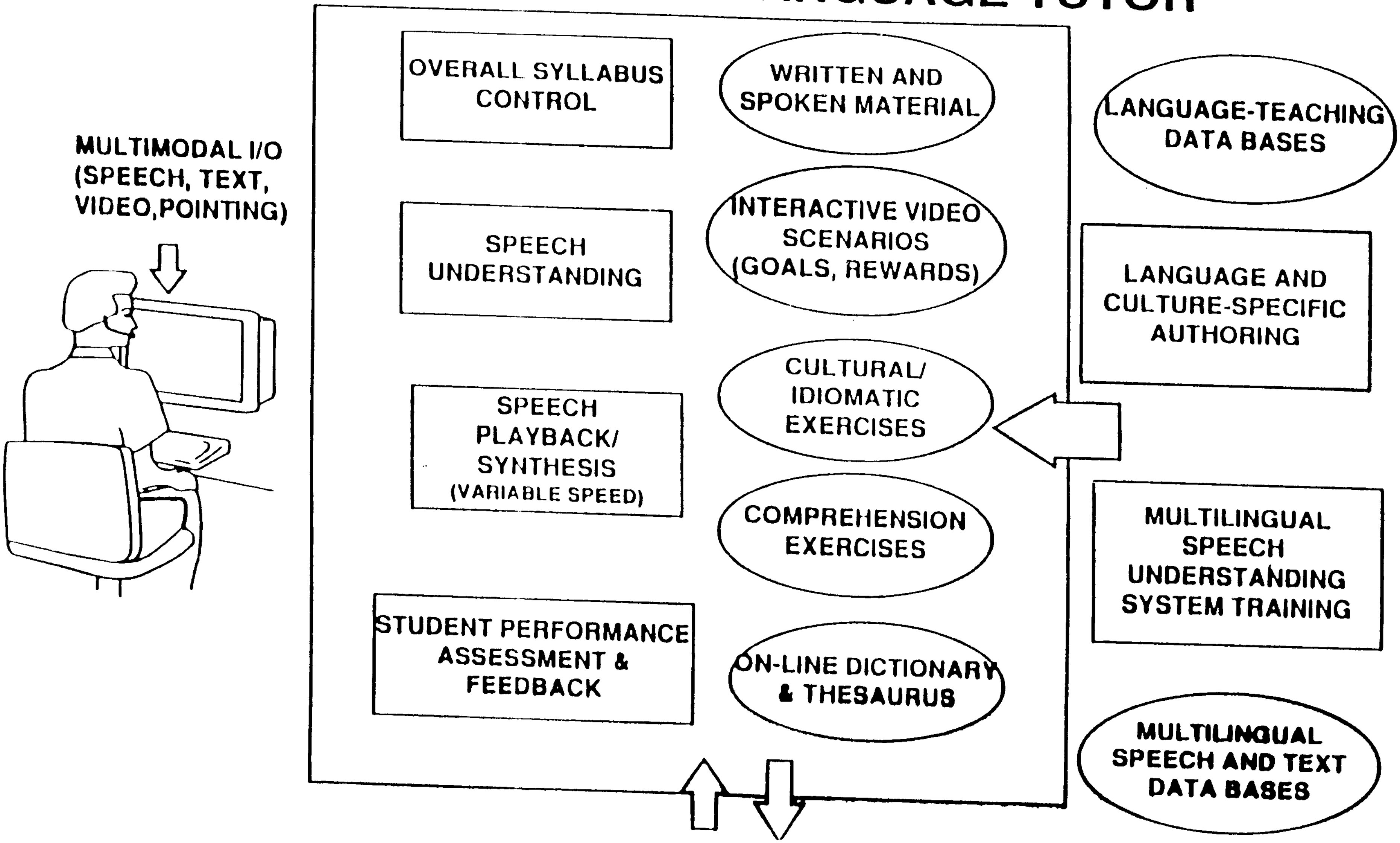
location

**_PARAPHRASES: ENGLISH AND FRENCH_**

| | | | | | |
|---|---|---|---|---|---|
| commander | of first | battalion | report | your | location |
| le commandant | du premier | bataillon | signalez | votre | emplacement |

첫째 대대의 부대장 네 장소를 보고하라

# INTELLIGENT FOREIGN LANGUAGE TUTOR



MULTIMODAL I/O (SPEECH, TEXT, VIDEO,POINTING)

OVERALL SYLLABUS CONTROL

SPEECH UNDERSTANDING

SPEECH PLAYBACK/ SYNTHESIS (VARIABLE SPEED)

STUDENT PERFORMANCE ASSESSMENT & FEEDBACK

WRITTEN AND SPOKEN MATERIAL

INTERACTIVE VIDEO SCENARIOS (GOALS, REWARDS)

CULTURAL/ IDIOMATIC EXERCISES

COMPREHENSION EXERCISES

ON-LINE DICTIONARY & THESAURUS

LANGUAGE-TEACHING DATA BASES

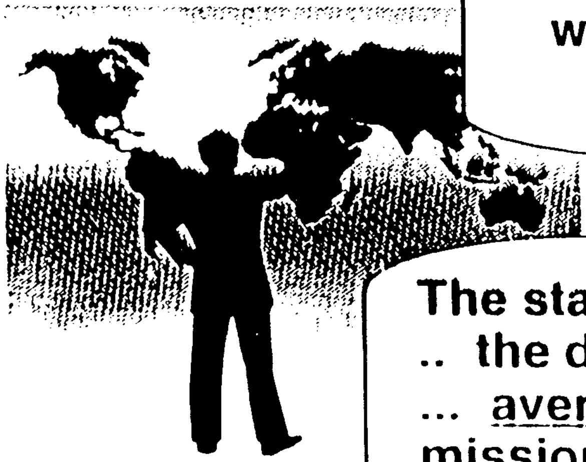
LANGUAGE AND CULTURE-SPECIFIC AUTHORING

MULTILINGUAL SPEECH UNDERSTANDING SYSTEM TRAINING

MULTILINGUAL SPEECH AND TEXT DATA BASES

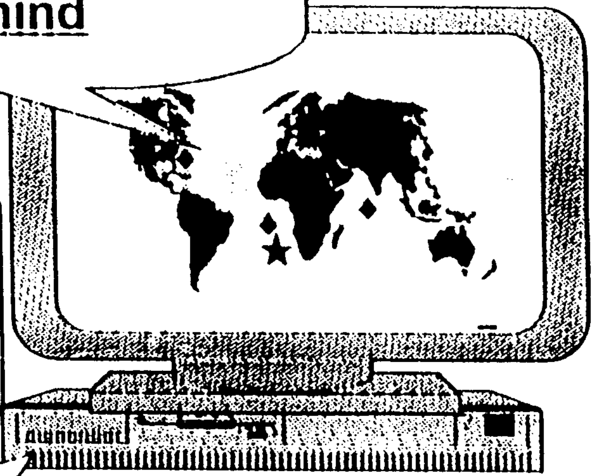
# Interactive decision support using dialog

## Impact: Improved military readiness, affordability, and usability



Show me the current position of the missiles
we shipped to the the middle east
...and show me where we expected them to be by now

The stars show the current in-transit positions
.. the diamonds show where they were planned to be
... average shipment is behind by 18 hours and the
mission critical shipment is 24 hours behind

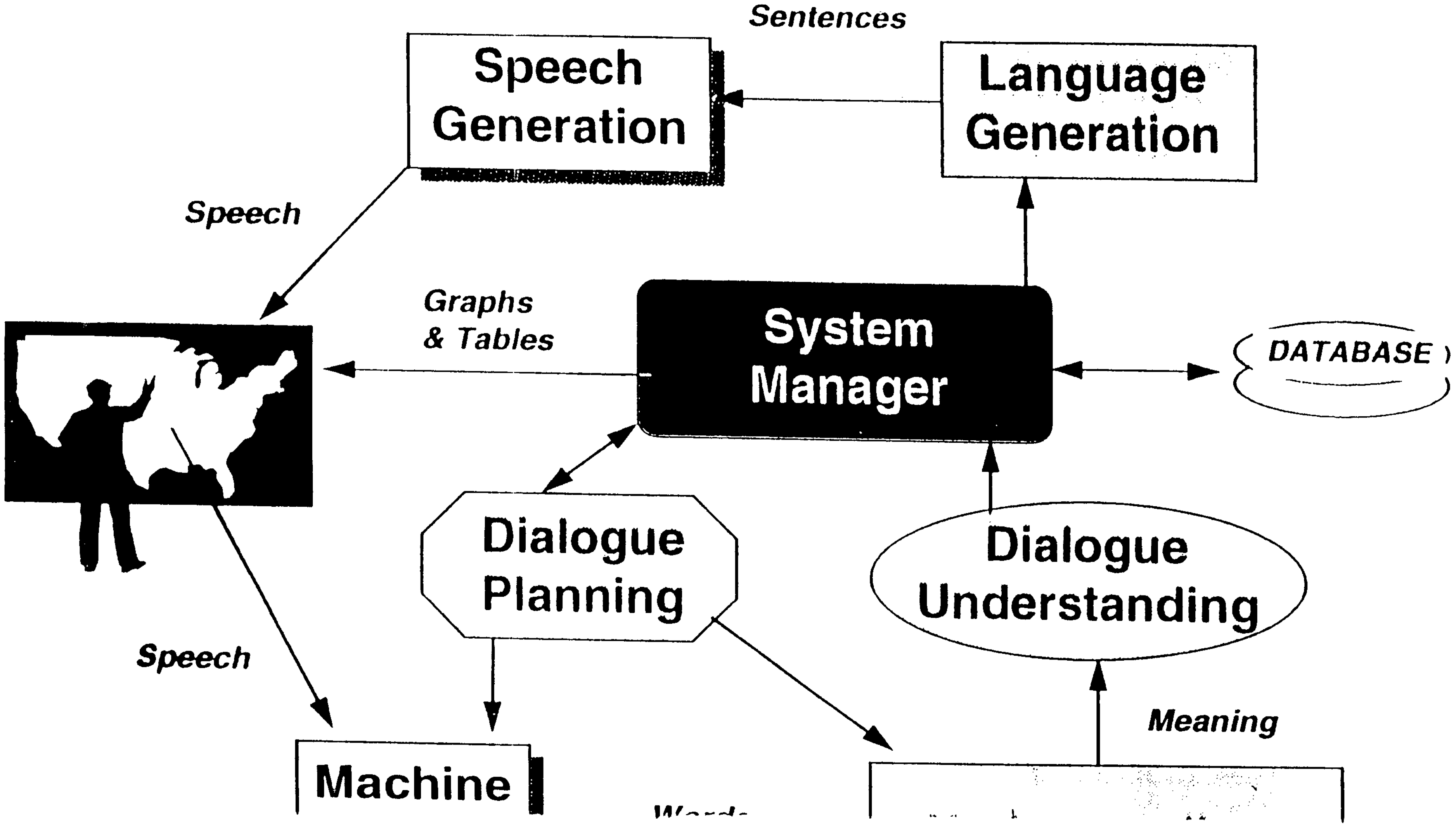


Set-up an immediate collaboration conference ..
.. include both transportation and logistic anchor desks

.. Oh .. and also show me the current warehouse
status of any remaining missiles plus seeker heads

Roger: Logistics and Transportation Anchors are set-up
.. 2-way video will cost the standard rate ..
.. there will be a 2 minute wait for the warehouse info.

# Dialogue System Architecture



# SPOKEN LANGUAGE TECHNOLOGY AND APPLICATIONS: STATE-OF-THE-ART

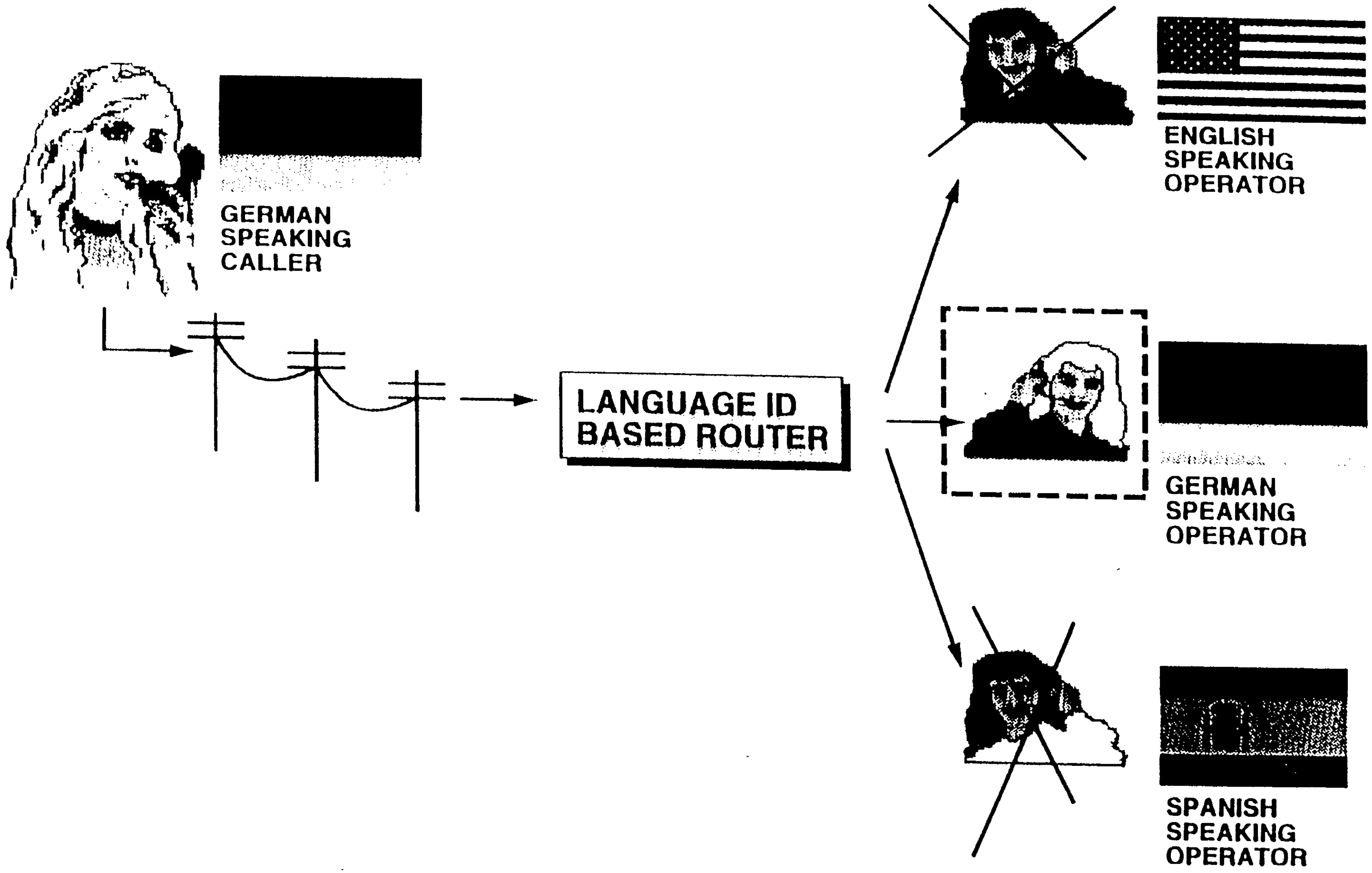
- VISION AND CHALLENGE
- TECHNOLOGY ISSUES AND PROGRESS
- MILITARY AND GOVERNMENT APPLICATIONS
- ➡ CURRENT THRUSTS IN DIALOGUE SYSTEMS, LANGUAGE EDUCATION, AND SPEECH TRANSLATION
- PLANS AND THE FUTURE

UNCLASSIFIED

## Examples of Users and Applications

| Users | Data Entry & Commun. | Data Access | Command & Control | Training | Intelligent Info. Retrieval | Translation |
|---|---|---|---|---|---|---|
| Soldier | xx | x | x | x |  | x |
| Naval CIC | xx | xx | xx | xx | x |  |
| Pilot | xx | x | xx |  |  |  |
| Agent | xx | xx |  | x | x | x |
| ATC | x | xx |  | x |  |  |
| Info. Analyst | x | x |  |  | xx | xx |
| Diplomat | x | x |  |  | xx | xx |
| Joint Force |  | xx | xx |  | x | xx |

# POTENTIAL APPLICATION: FRONT-END ROUTER TO A BANK OF DIRECTORY ASSISTANCE OR 911 OPERATORS



# COMMON LANGUAGE VOICE RECOGNITION TRANSLATOR

(DESCRIPTION OF REQUIRED TECHNOLOGY FROM SENIOR WORKING GROUP REPORT ON MILITARY OPERATIONS OTHER THAN WAR, MAY 1994)

- **DESIRED CAPABILITIES**
  - **REAL-TIME VOICE TRANSLATION, ENGLISH TO FOREIGN LANGUAGE AND VICE VERSA**
  - LANGUAGE PRIORITY ON BASIS OF LIKELIHOOD OF U.S. INVOLVEMENT
- **RATIONALE**
  - COALITION WARFARE AND OOTW
  - WORLDWIDE MILITARY INVOLVEMENT IN OOTW
  - U.S. LACK OF RANGE OF LINGUISTS IN MILITARY FORCES
  - BENEFITS: MORE EFFECTIVE COMMUNICATION, INTELLIGENCE, TRAINING
- **OPERATIONAL CONCEPT**
  - RECOGNIZE, UNDERSTAND, TRANSLATE, VERIFY
  - EMPLOY COMMON COALITION LANGUAGE FOR TRANSMISSION AND PORTABILITY TO MULTIPLE LANGUAGES
- **APPLICABILITY**
  - MILITARY, DIPLOMATIC, LAW ENFORCEMENT
  - COMMERCIAL
- **RELATED TECHNOLOGY AREAS**
  - SPEECH RECOGNITION, UNDE

UNITED STATES SPECIAL OPERATIONS COMMAND

# Technology Review: Special Operations Forces (SOF) Speech Recognition for Language Sustainment

## Descriptions of Speech Recognition Systems

AIRBORNE

ARMY RESEARCH INSTITUTE

## TriplePlay *Plus!*

Dr. Martin Rothenberg

TriplePlay *Plus!*, from Syracuse Language Systems, is a fund and effective way to learn to read, speak, and understand a foreign language. The unique *Speech Recognition* mode in TriplePlay *Plus!* bring language learning closer to the natural way a person learns a first language -- by spoken interaction.

TriplePlay *Plus!* features *Speech Recognition* technology licensed from Dragon Systems, Inc., that evaluates the learner's pronunciation. *Speech Recognition* is embedded in interactive games and conversations that provide an engaging multimedia-immersion approach to language learning.

TriplePlay *Plus!* includes a high-quality dynamic microphone for use with the *Speech Recognition* and record/Playback features. The Windows CD-ROM is co-published by Syracuse Language Systems, Inc., and Random House, Inc. as part of the Living Language Multimedia™ product line.

Designed for learners age 8 to adult, TriplePlay *Plus!* teaches over 1,000 words and phrases in versions for learning Spanish, French, German, English or Hebrew. The produce uses *multimedia language immersion*, a learning method developed at Syracuse University, to teach naturally, entirely in the language to be learned.

TriplePlay *Plus!* is the winner of several industry awards, including a 1995 *HOME PC* Editor's Choice Award, a 1994-1995 *Technology & Learning* Award of Excellence, and a 1994 *NewMedia* INVISION Award for innovation in multimedia.

Contact: Dr. Martin Rothenberg
Syrcause Language Systems, Inc.
719 E. Genesee St.
Syracuse, NY 13210
(315) 478-6729/(800) 688-1937; FAX: (315) 478-6902

# Conversim™--A Dialog with a Native Speaker in a Multimedia Environment

Dr. William G. Harless

Through the creative application of interactive video and speech recognition technologies, Interactive Drama's Conversim software offers a unique approach to foreign language training: Students learn to speak the language through face-to-face dialogue with native speakers in simulated real-life situations.

Two simulations will be presented: "Medical Spanish" and "Roberto's Restaurant." The simulated character in the medical Spanish program is an elderly real patient with a history of heart trouble. The simulated character in the restaurant program is actually the charismatic owner of the restaurant. Each simulation involves a situation which requires that students master words and phrases in order to manage the real-life situation. Assisted by an on-screen native instructor, students first learn and rehearse the vocabulary, then they practice using this vocabulary in a direct dialogue with the simulated character.

Contact: Dr. William G. Harless
Interactive Drama, Inc.
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 20814
(301) 654-0676; FAX: (301) 657-9174
e-mail: INTDRAMA@aol.com

## The Here and Now in Voice-Interactive Language Learning Systems

LTC Steve LaRocca and COL Woody Held

In developing voice-interactive systems for foreign language study at West Point, speech recognition was added as an enhancement to interactive video platforms. The idea was to make existing language lessons "talkies" by using speech recognition in lieu of a keyboard or mouse to respond to multiple choice questions. The speech recognition technology used is inexpensive and relatively simple. The recognizer is used to differentiate between a small number of complete utterances, trained specifically for each lesson. This system adds vocabulary development to the work of authoring lessons, yet provides students with courseware that uses all four languages skills (listening, reading, writing and speaking) and more realism as well. Voice-interactive systems at West Point capitalize on the low cost ($150) of Prometheus Aria 16SE sound cards and the easy-to-use Aria Listener software. We are working with Duke University to bring Aria-type speech recognition into the WinCALIS authoring system.

Contact: LTC Steve LaRocca
Center for Technology Enhanced Language Learning
Department of Foreign Languages
U.S. Military Academy
West Point, NY 10996
(914) 938-5286; FAX: (914) 938-3585
e-mail: gs0416@usma3.usma.edu

## Speech and Language Technology

Dr. Madeleine Bates and Mr. Sean Colbath

We will demonstrate or show on videotape a number of systems that illustrate the state of the art in speech recognition and language understanding:

1. ATIS - an air travel information system that understands spoken questions and commands.

2. Large vocabulary (20,000 words), real-time, continuous, speaker-independent speech recognition.

3. Form filling via speech.

4. Speaker identification - identifies which speaker form a known set of possible speakers is talking, very rapid enrollment process, works in any language.

5. VALAD - a system that integrates speech with mouse, menus, and keyboard, interfacing to the logistics anchor desk and intended for use by military logistical planners. The resulting interactive spoken language understanding system was recently demonstrated at Prairie Warrior '95.

Contact: Dr. Madeleine Bates
BBN Systems and Technologies
70 Fawcett Street
Cambridge, MA 02138
(617) 873-3634; FAX: (617) 547-8918
e-mail: Bates@BBN.com

## Language Tutor and Bilingual Voyager System

Dr. Victor W. Zue, Dr. Joseph Polifroni, and Dr. Stephanie Seneff
Spoken Language Systems Group

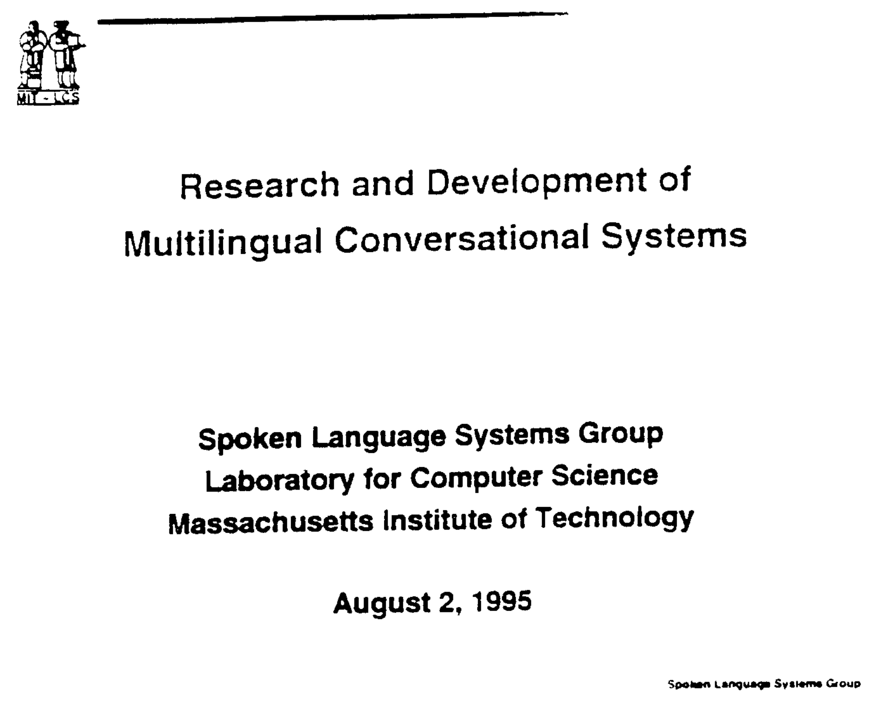
The Spoken Language systems Group will demonstrate two related systems:

1. A "Language Tutor" applied to Japanese, which provides users with practice drills and feedback to help them recall and pronounce Japanese words and phrases that will be of use in the second demo.

2. The "Bilingual Voyager" system, which gives the user information appropriate for a traveler in Cambridge, Massachusetts (hotels, restaurants, banks, etc.) and locates places of interest on the map. The user can converse with the system in English, Japanese, or "mixed mode" (e.g., user speaks in English, system responds in Japanese).

Both systems use a continuous-speech, speaker-independent speech recognizer. The acoustic models were trained on both read and spontaneous speech from native speakers in each language. The systems run on a Sun Sparc 20 workstation.

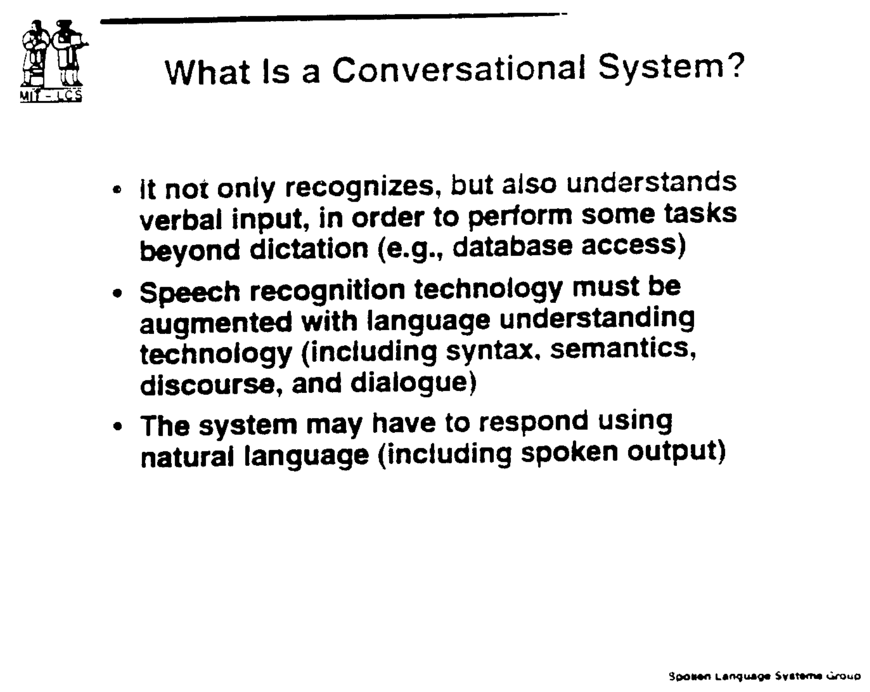
Contact: Dr. Stephanie Seneff
Spoken Language Systems, Group
Massachusetts Institute of Technology
Cambridge, MA 02139
(617) 253-0451; (FAX): (617) 258-8642
e-mail: seneff@lcs.mit.edu



# Research and Development of Multilingual Conversational Systems

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology

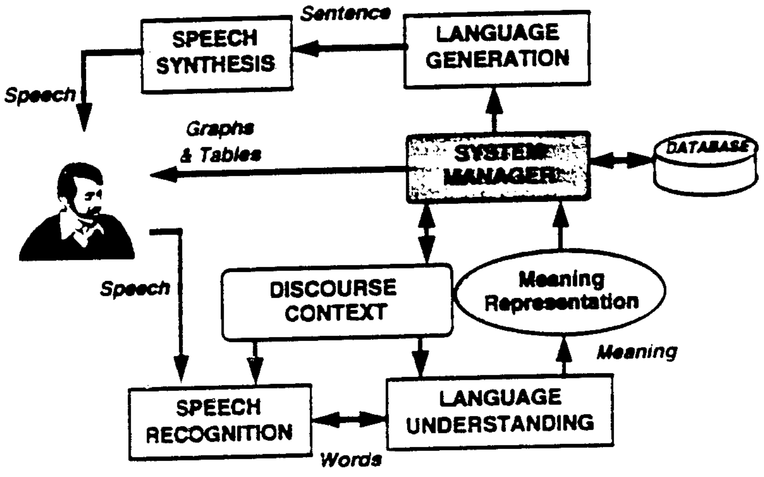
August 2, 1995



## What Is a Conversational System?

- It not only recognizes, but also understands verbal input, in order to perform some tasks beyond dictation (e.g., database access)
- Speech recognition technology must be augmented with language understanding technology (including syntax, semantics, discourse, and dialogue)
- The system may have to respond using natural language (including spoken output)

## Conversational System Architecture



SPEECH SYNTHESIS

Sentence

LANGUAGE GENERATION

Speech

Graphs & Tables

SYSTEM MANAGER

DATABASE

Speech

DISCOURSE CONTEXT

Meaning Representation

Meaning

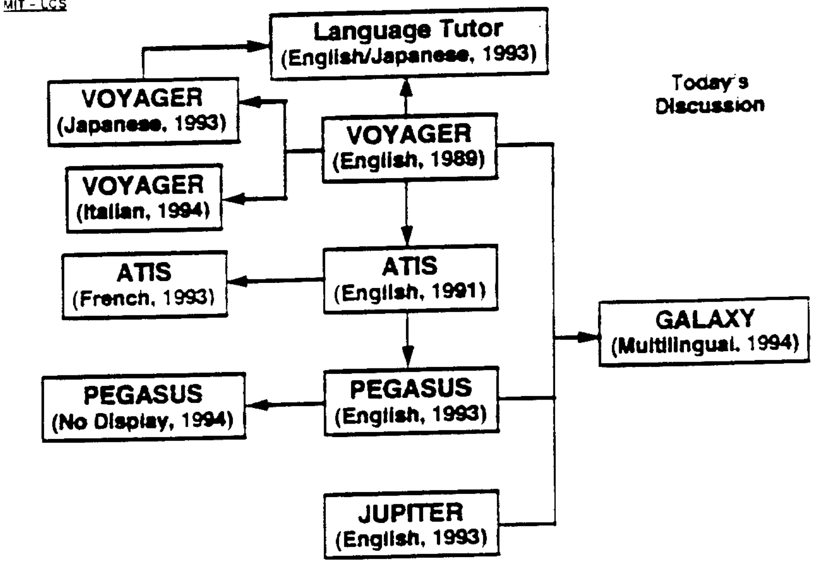
SPEECH RECOGNITION

LANGUAGE UNDERSTANDING

Words

Spoken Language Systems Group

## History of System Development at MIT



Language Tutor (English/Japanese, 1993)

Today's Discussion

VOYAGER (Japanese, 1993)

VOYAGER (English, 1989)

VOYAGER (Italian, 1994)

ATIS (French, 1993)

ATIS (English, 1991)

GALAXY (Multilingual, 1994)

PEGASUS (No Display, 1994)

PEGASUS (English, 1993)

JUPITER (English, 1993)

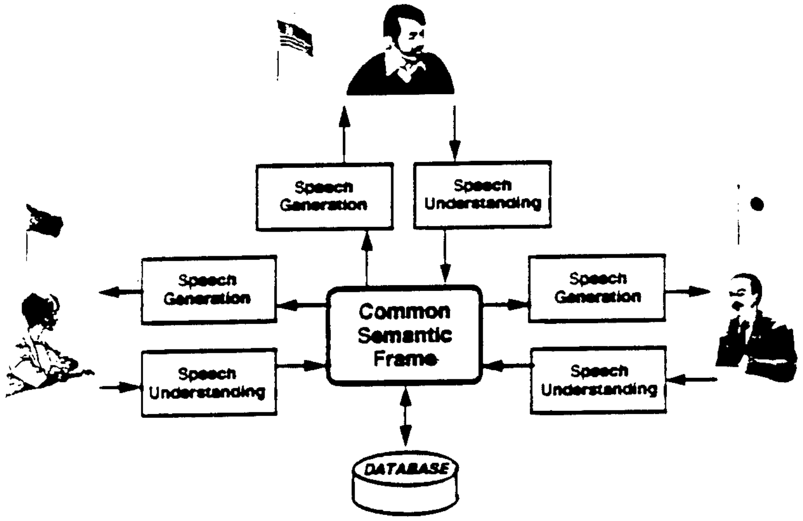
Spoken Language Systems Group

## Current Status at MIT

**Conversational systems are emerging that can:**

- **Deal with continuous speech, by unknown users, drawn from a large vocabulary,**
- **Understand the meaning of the utterances and take appropriate actions,**
- **Operate in real (or realistic) domains,**
- **Handle multiple languages (English, Japanese, Spanish, French, Italian, German, Chinese), and**
- **Deliver these capabilities in real-time, using standard workstations with no additional hardware**

## Multilingual Conversational Systems for Human-Computer Interactions



Speech Generation

Speech Understanding

Speech Generation

Speech Understanding

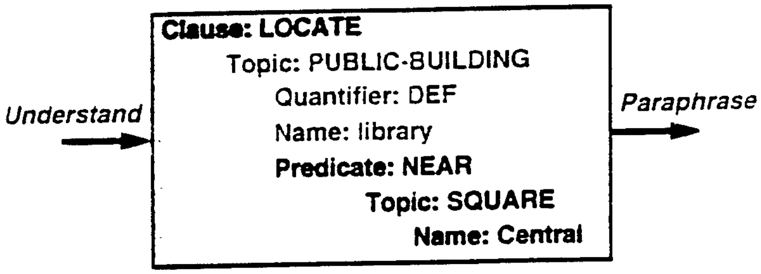
Common Semantic Frame

Speech Generation

Speech Understanding

DATABASE

## Semantic Frame Representation



Understand

**Clause: LOCATE**
Topic: PUBLIC-BUILDING
Quantifier: DEF
Name: library
**Predicate: NEAR**
**Topic: SQUARE**
**Name: Central**

Paraphrase

**WHERE IS** THE LIBRARY **NEAR CENTRAL SQUARE**

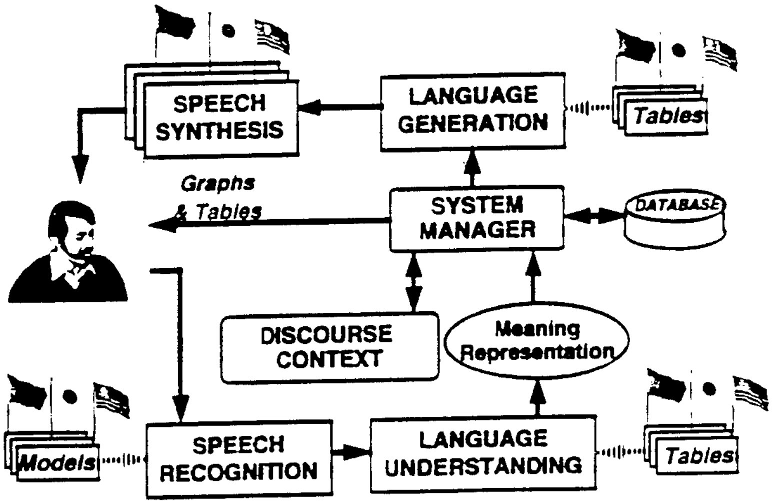
**SENTORARU SUKUEA NO CHIKAKU NO** TOSHOKAN WA **DOKO DESU KA**

**DOVE STA** LA BIBLIOTECA **VICINO A CENTRAL SQUARE**

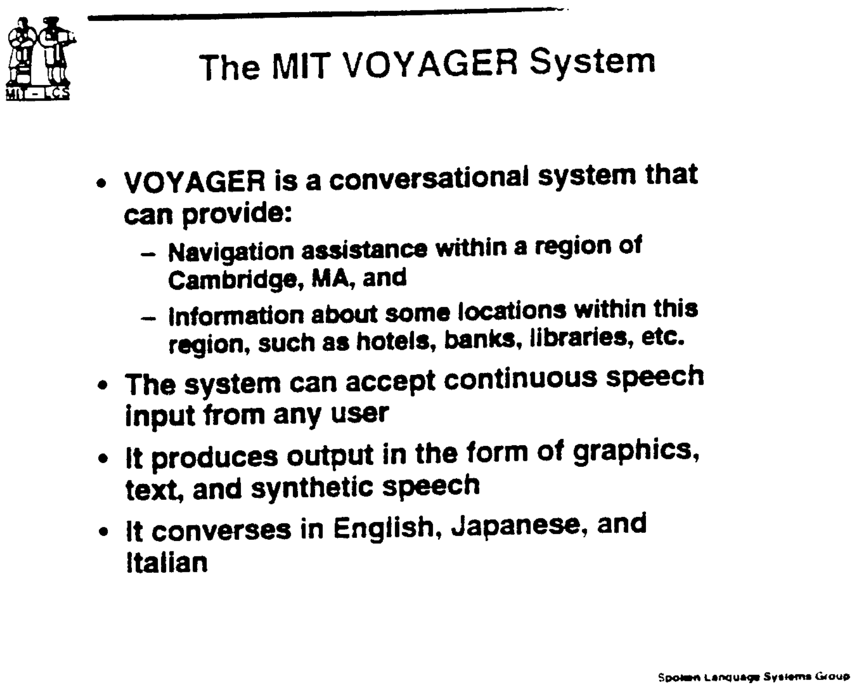
**OÙ SE TROUVE** LA BIBLIOTHEQUE **QUI EST PRÈS DE CENTRAL SQUARE**

Spoken Language Systems Group

## Multilingual Conversational System



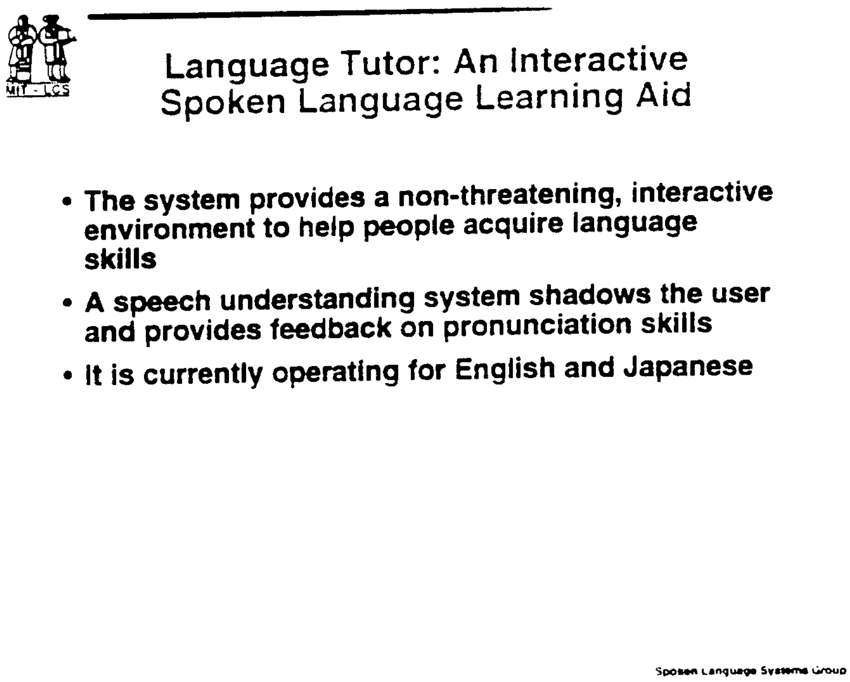
Spoken Language Systems Group



## The MIT VOYAGER System

- VOYAGER is a conversational system that can provide:
  - Navigation assistance within a region of Cambridge, MA, and
  - Information about some locations within this region, such as hotels, banks, libraries, etc.
- The system can accept continuous speech input from any user
- It produces output in the form of graphics, text, and synthetic speech
- It converses in English, Japanese, and Italian

Spoken Language Systems Group

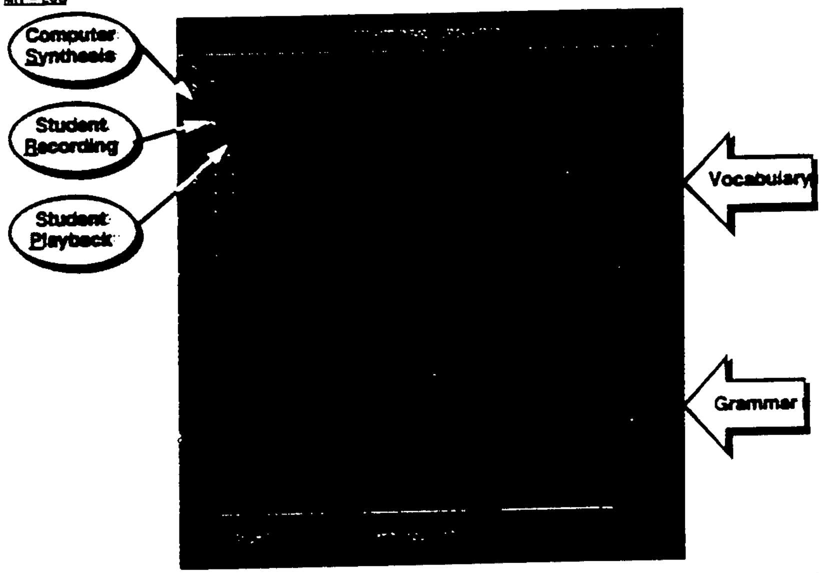


## Language Tutor: An Interactive Spoken Language Learning Aid

- The system provides a non-threatening, interactive environment to help people acquire language skills
- A speech understanding system shadows the user and provides feedback on pronunciation skills
- It is currently operating for English and Japanese

Spoken Language Systems Group

## Language Tutor Display



Spoken Language Systems Group

## A Novel Approach to Language Learning

- Dovetails a language tutor with a multilingual conversational system such as VOYAGER
- Each lesson would consist of:
  - Newly introduced vocabulary and grammar drills
  - A scenario specifically designed for the lesson
- Students can speak in their native language and hear responses in target language, or vice versa, providing flexible alternatives for practicing speaking/listening
- Enables students to practice interaction in a risk-free setting
  - Goes beyond mechanics of standard reading/speaking exercises.
  - Simulates real world in a language laboratory.

Spoken Language Systems Group

# Voice Interactive Language Training System (VILTS)

Patti Price, Marikka Rypa, Leo Neumeyer, and George Chen
Research and Technology Laboratory
SRI International

Mike Valatka and Kathleen Egan
Office of Research and Development, CIA

Helena Hughes
Federal Language Training Laboratory, CIA

Jacqueline Pogany
Office of Training and Education, CIA

## 1.0 Overview

The Voice Interactive Language Training System (VILTS) is language education software being developed to foster improvement in French comprehension and speaking skills. VILTS represents a joint development effort between SRI International, the Office of Training and Education (OTE), and the Federal Language Training Laboratory (FLTL). The focus of the program is to train students at levels 1 through 3 in comprehension and discrimination skills and subsequently in speaking and pronunciation skills through a series of activities centered around listening, speaking, and reading. SRI is incorporating advances in its research in speech recognition and pronunciation evaluation to provide students with the opportunity to navigate through a unit using oral communication, with the system recognizing appropriate or inappropriate responses. At the end of a unit, the student will be given feedback as to how s/he compares to a native speaker, and additional feedback on specific problematic sounds. Pronunciation exercises will be provided that target specific problem areas tailored to specific student needs.

The present system under demonstration uses French speech recognition capabilities; the evaluation capabilities are scheduled to be included in early 1996.

## 2.0 Speech Recognition and Speech Evaluation

As a leader in speech technology, SRI has conducted world-class research in speech recognition, pronunciation evaluation, and speech processing capabilities as applied to language education. SRI has consistently scored among the top contenders in the ARPA-sponsored speech benchmark competitions in the last 10 years; SRI's speaker-independent technology can recognize natural, continuous speech without requiring the user to train on the system. The VILTS represents a pioneering effort to combine the power and robustness of state-of-the-art speech recognition with pedagogically engaging learning activities and feedback on individual pronunciation.

### 2.1 Speech Recognition Activities

The student interacts with the system orally to simulate natural conversation by responding to questions or posing questions to the system.

As student speech is elicited through a variety of activities, the French speech recognizer listens for the oral student input and responds appropriately, either accepting or rejecting the response, depending on a threshold level of acceptance. The extent to which the student or instructor can determine this level of acceptance is an area of future investigation.

## 2.2 Speech Evaluation

As the student completes a unit and enough speech has been collected, pronunciation evaluation algorithms will be employed to compare the student performance level to the pronunciation of a native speaker. Ratings from expert French instructors are being collected as part of this development, and the ratings by machine will correlate with the expert raters. As a result of evaluation scores and subscores, the system will suggest and provide exercises to improve a student's problem areas.

## 3.0 Pedagogical Architecture

The design and development of the Voice Interactive Language Training System represents a collaboration between SRI International, the Office of Training and Education, and the Federal Language Training Laboratory. The units and activities are being developed by instructional design professionals at all three institutions; FLTL is developing the graphics which are being integrated into the program by SRI.

Using spontaneous, unscripted French conversations on various topics and excerpts from the French newspaper LeMonde, the VILTS program provides the student with authentic, unrehearsed French speech as might be heard in everyday speech in France. The conversations are the basis for the activities, which focus on comprehension, speech production, and pronunciation. These units can be used to complement/supplement a course for students learning French, or they can be used to support maintenance training, self-study, and refresher programs.

Conversations on ten different topics, including such areas as travel, health care, education, and politics were collected from a pool of 100 native speakers of French. A read version of these conversations was subsequently recorded by the same speakers so that both spontaneous speech and a clearer and slower version is available to the student. Conversations were collected to approximate three distinct levels of student ability; beginning, intermediate and advanced, corresponding roughly to government standard levels 1, 2, and 3. The student chooses a level of conversation with which to work, and then chooses from a menu of topics available at that level. Each lesson contains activities centering on listening, speaking, and reading.

Contact: Dr. Patti Price
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
(415) 859-5845; FAX: (415) 859-5984
e-mail: pprice@speech.sri.com

# Project LISTEN:
# A Reading Coach That Listens

Dr. J. Mostow, Dr. M. Eskenazi, Dr. A. Hauptmann,
Dr. B. Milnes, and Dr. S. Roth

Project LISTEN is developing a novel weapon against illiteracy -- an automated reading coach that displays a story on a computer screen, listens to a student read it aloud, and helps where needed. The coach provides a combination of reading and listening, in which the student reads wherever possible, and the coach helps wherever necessary. The coach was demonstrated at ARPA's 1994 Human Language Technology Workshop, featured in BYTE's cover story on "7 New Ways to Learn," and honored with the Outstanding Paper Award at the 1994 National Conference on Artificial Intelligence.

Problem: Literacy is essential to economic and military effectiveness in the Information Age. For example, both industry and military need a pool of recruits who can read and comprehend manuals for high-tech equipment. Illiteracy costs the United States over $225 billion dollars annually in corporate retraining, lost competitiveness, and industrial accidents. People with low reading proficiency are often unemployed, poor, or incarcerated. A reading coach that listens could give millions of American children and adults individualized reading assistance that teachers cannot provide.

Approach: Project LISTEN exploits an opportunity created by advances in speech technology, reading, and human-computer interaction. The reading coach adapts Carnegie Mellon's state-of-the art Sphinx-II speech recognizer to analyze the student's oral reading. The coach responds with assistance modelled after expert reading teachers. Successive prototypes have been tested on approximately 100 children in Pittsburgh public schools. To go from prototype to practice, the coach must be deployed in schools, evaluated in actual use, and refined into a practical educational tool.

Impact: Project LISTEN offers a powerful new tool to combat the literacy crisis that threatens the nation's economic and military security. Second, as one of the first "stress tests" of real-time continuous speech recognition in a real application, Project LISTEN provides valuable technical lessons about how to make spoken communication with computers usable and robust. Finally, applications to defense needs include more cost-effective reading instruction for the 95,000 children enrolled in Department of Defense Dependents Schools. Spinoff applications include individualized foreign language training for Special Forces personnel.

Contact: Dr. Jack Mostow, Director
Project LISTEN
Carnegie Mellon University Robotics Institute
215 Cyert Hall, 4910 Forbes Avenue
Pittsburgh, PA 15213-3890
(412) 268-1330; FAX: (412) 268-6298
Internet: mostow@cs.cmu.edu

# Entropic Speech Technology in Language Education.

Dr. Jared Bernstein

Entropic Research Laboratory has formed a Language Systems Group to develop algorithms and build products for language instruction and evaluation. Entropic's existing Unix software products provide the base technology for Interactive Spoken Language Education (ISLE). Entropic offers systems and tools to support high-accuracy speech recognition for large vocabularies, and for manipulation, storage and synthesis of high-quality speech. Entropic's core products are advanced signal processing software and virtual instruments for the research and development community. Over 400 R&D groups conduct their research and build products with Entropic tools.

Fluency Demonstration System (English): Spoken English can be aligned with corresponding text and used to automatically judge the speaker's fluency.

Picture Demonstration System (English/Spanish): An example of robust, tolerant speech recognition in a multiple choice exercise.

Animation Demonstration System (English/Spanish): An example of interaction in Spanish or English to control animated events.

Entropic Time Scale Modification (language independent): Software that slows down or speeds up recorded speech without distortion.

The following pages describe the Entropic program.

Contact: Dr. Jared Bernstein
Language Systems Group
Entropic Research Laboratory, Inc.
1040 Noel Drive
Menlo Park, CA 94025
(415) 328-8877; FAX: (415) 328-8866
e-mail: jared@entropic.com

# Foreign Language Dialog System

Dr. John T. Lynch and Dr. Beth Carlson

The FOREIGN LANGUAGE DIALOG SYSTEM is a speech recognition-based automated tool for providing a novice language learner with authentic practice in speaking and listening to a second language. The tool also can provide a convenient way to maintain one's language skills. We have developed a proof-of-concept demonstration system using UNIX-based research software in order to illustrate the potential for providing an environment where a learner can focus on the immediate communication task as opposed to a memorization or drill exercise. The DIALOG SYSTEM therefore complements foreign language instruction whether it involve machine or human interaction. The DIALOG SYSTEM would ideally be integrated with other instruction so that the vocabulary and grammar of the DIALOG SYSTEM would match the requirements of the learner at a particular stage of progress. In addition, the content of the DIALOG SYSTEM's scenarios could be matched to the specific needs of the learner, e.g., food distribution, heath care, or combat operations.

The DIALOG SYSTEM is designed with the following three principles in mind.

1. To engage the learner more fully, the learner's speech should determine the system response.

2. To be realistic, the exact wording (vocabulary, grammar) should be open and not constrained by the system.

3. To improve the accuracy of the speech recognition system so that the system is useful, the intention and meaning of the learners utterances should be constrained. This can be done by context defined by the scenario.

Our system addresses these principles by having the learner address verbal questions to a person represented on the screen. Our present system uses clip-art images but future versions would use photographs or motion video of native speakers which would further enhance the immersion experience.

The demonstration system is based on a security interview scenario. To help guide the learner, the system provides a form to be filled out for the subject who is being interviewed. This form would specify an issue such as "foreign travel" but would leave unspecified how the learner would elicit the necessary information from the subject being interviewed. That is, the system would respond to a variety of wordings (expected of the learner at a specified level of language achievement). To further aid the learning process, the system can also provide suggestions on how to formulate each question, if the user requests such information. Other scenarios are easily envisioned: admission to a hospital, interrogation of a suspected spy, ordering and planning distribution of food supplies. We plan to provide tools so that language instructors can easily develop scenarios matched to the needs of their training programs.

The current proof-of-concept system is implemented in three languages: English, Spanish, and German.

The English system has two characters to be interviewed and they can each be asked 25 questions in a variety of wordings. For example, one can ask: "Have you been overseas recently?", or "Any overseas travel in the last 3 years?" The English speech recognition system was trained on a general speech corpus called TIMIT which consists of about 4 hours of studio-quality phonetically rich speech.

The Spanish system has two characters who can be asked five questions each with a number of varients per question type. The Spanish speech recognizer was trained on data collected from 8 male and 8 female talkers who varied from native speakers to experienced learners to novice speakers. The German system has one character who can be asked five questions (with two wordings each).

The German recognizer was trained on data collected from 3 males and 3 females who are novice to medium experienced speakers. Our long term plans include providing tools so that language instructors could port the system to new languages by collecting appropriate data and training new speech models. While training data collection is not always desirable, it is often necessary for less common languages for which suitable data is not easily obtained.

The system demonstrated can run in real-time on both a SPARC 10 UNIX workstation and a 486/Pentium-based personal computer running the LINUX operating system. The speech recognizer software is based on HTK (Hidden Markov Model Toolkit), which is commercially available through Entropic, and uses a continuous speech recognition algorithm with a language grammar. Modifications were made to the recognition algorithm to accept live speech input and to interact with the graphical user interface (GUI). The GUI is based on the MOTIF X-WINDOWS programming software. The current configuration of the system uses several research components that are combined through the use of data pipes and shell scripts. Future general system design improvements are needed to increase system and response speed and to improve the human machine interface. In addition, further enhancements to the actual speech recognizer include modeling the speech of talkers at various points along the novice to native continuum. The system could then be responsive to the level of a particular learner and at the same time provide level-specific pronunciation feedback to that learner.

Contact: Dr. John T. Lynch
MIT Lincoln Laboratory
244 Wood St. - Rm S4-177
Lexington, MA 02173-9108
(617) 981-2746; FAX: (617) 981-0186
e-mail: jtl@sst.ll.mit.edu

Technology Review:
Special Operations Forces (SOF)
Speech Recognition for Language Sustainment

# Appendix C:
# Speech and Text Translation Systems

# Technology Review:
# Special Operations Forces (SOF)
# Speech Recognition for Language Sustainment

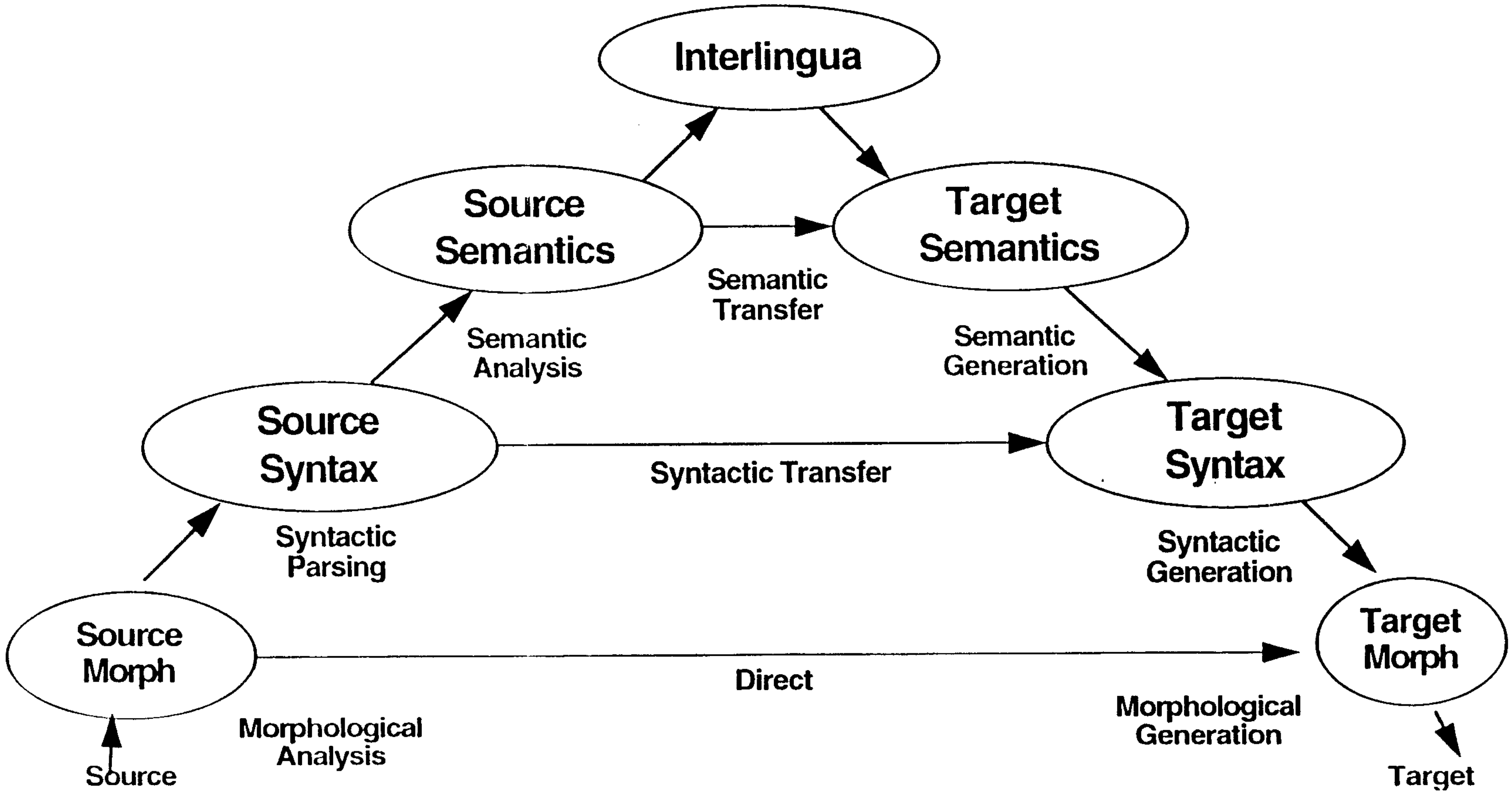
## Dr. Susann Luperfoy's Presentation
## "Voice-to-Voice Machine Translation: Problems and Prospects"

# Voice-to-Voice Machine Translation: Problems and Prospects

SOF Workshop on
Speech Recognition for Language Sustainment
3 August 1995

Susann Luperfoy
luperfoy@mitre.org

# Standard Classification of Machine Translation Systems



Interlingua

Source Semantics

Target Semantics

Semantic Transfer

Semantic Analysis

Semantic Generation

Source Syntax

Target Syntax

Syntactic Transfer

Syntactic Parsing

Syntactic Generation

Source Morph

Target Morph

Direct

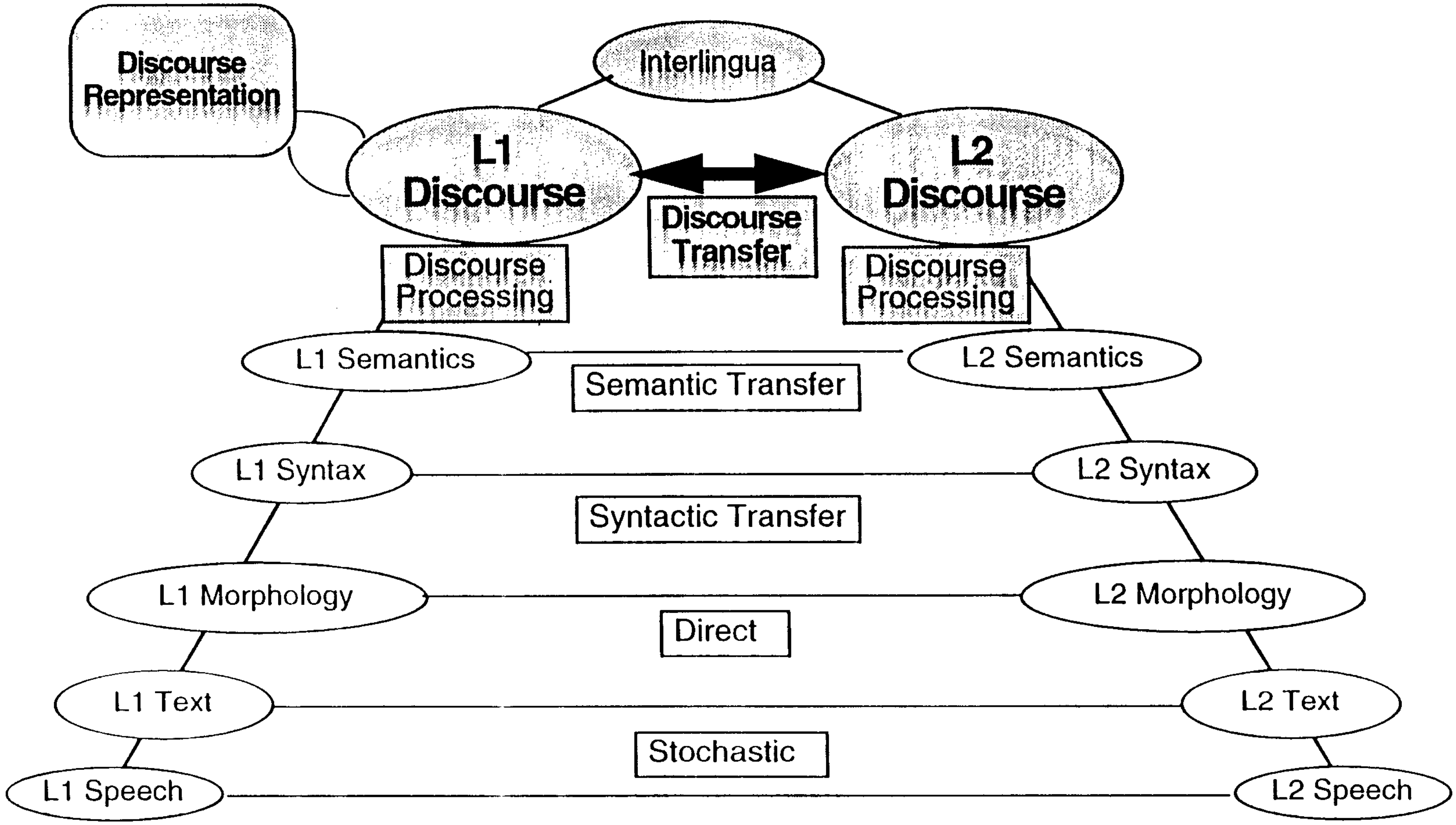
Morphological Analysis

Morphological Generation

Source

Target

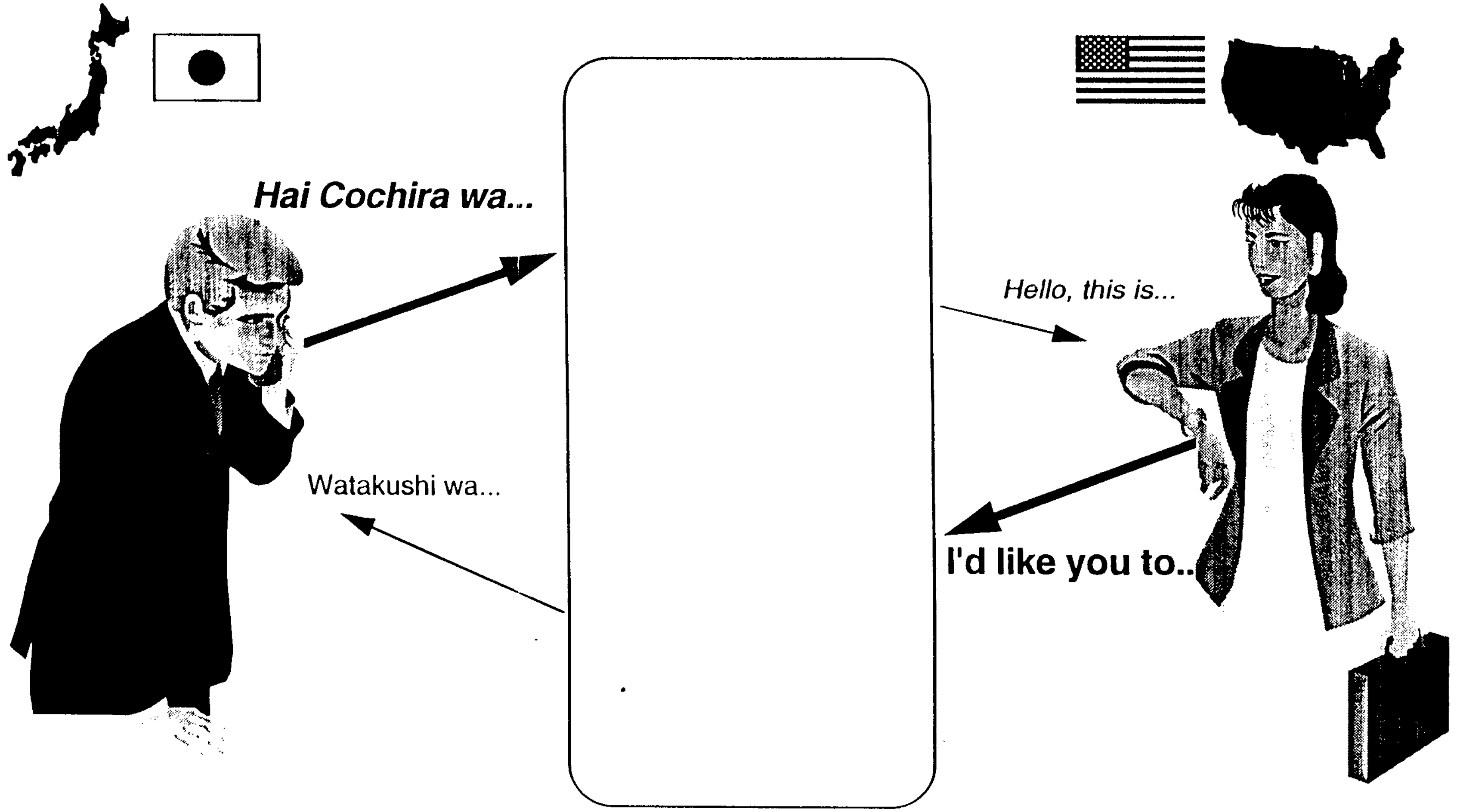
# Discourse for Interactive Speech-to-Speech Machine Translation



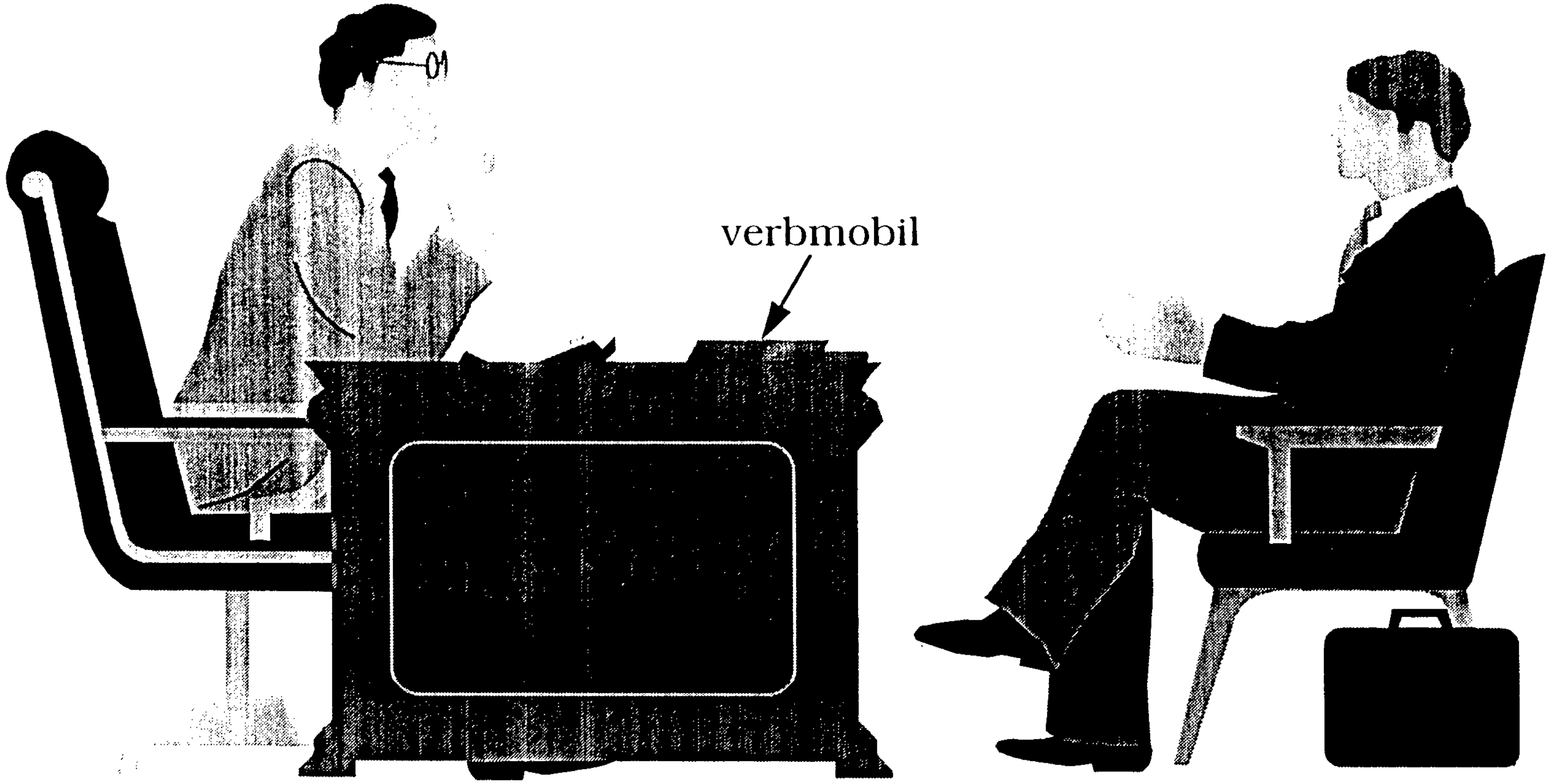
# Dialogue Processing

- Dialogue Manager :=
  - Discourse Processor + Interaction Manager
- Dialogue Tracking
  - Repair sequences
  - Must keep both sides of dialogue
- Two services from Discourse results
  - directly affect user interface dialogue behavior
  - indirectly by assisting other modules

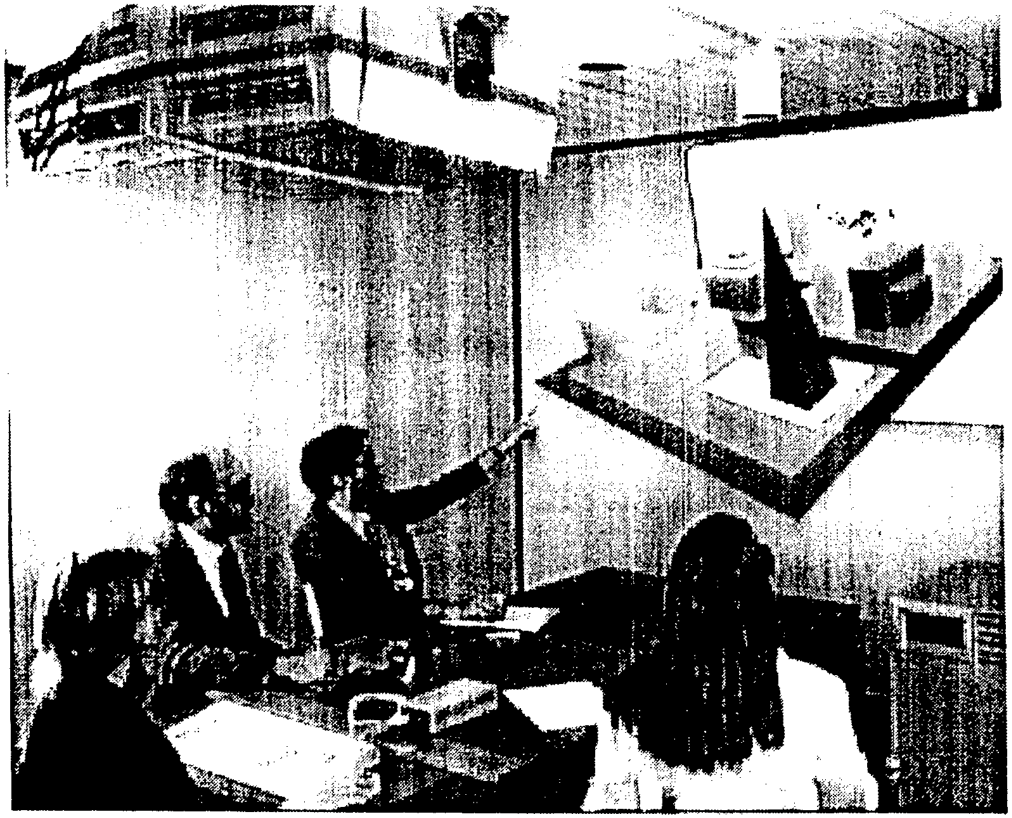
# Interpreting Telephone



# Verbmobil



# Potential Application of Verbmobil



- shared visual environment
- non-verbal communication
- dialogue management handled by users

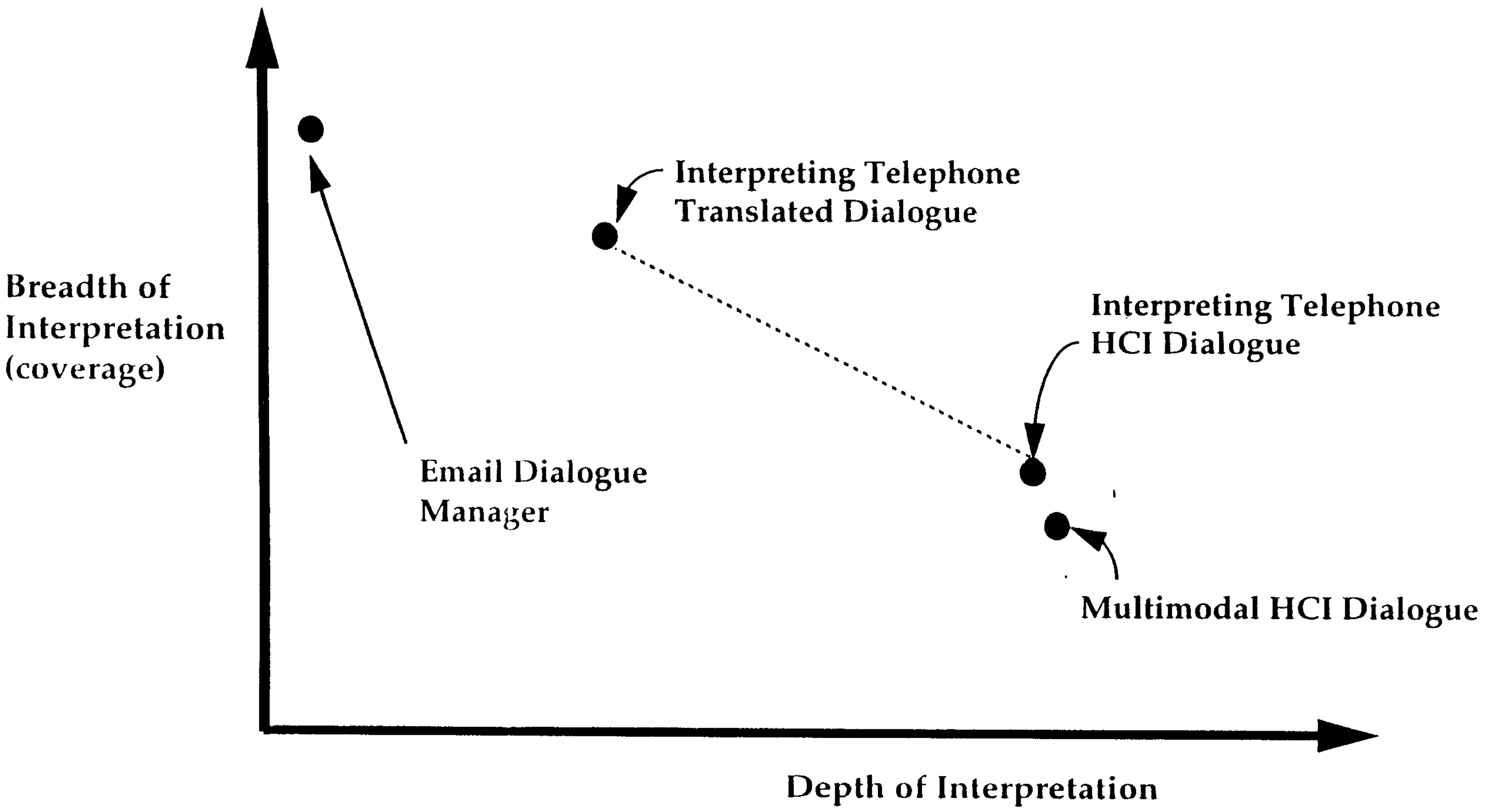
# Machine Aided Voice Translation (MAVT)

- Machine-Aided Voice Translation
- Users share visual context
- Non-verbal communication
- All verbal communication transmitted by system

# The Three Dialogue Types

- Device-Human
  - monolingual (untranslated)
  - full NLU
- Human-Human
  - intentions, task goals
  - bilingual
- Translator-Human
  - monolingual
  - partial NLU

# NL System Design Space



# Address Ellicitation in Samples

We'll send you a registration form.<br>
Your name and address, please?

**My address is 23 Chayamachi, Kita-ku, Osaka.**
**My name is Mayumi Suzuki.**

All right.
We'll send you a registration form.

Your name and address, please?

**My address is 2-2 Tokui-machi, Higashi-ku, Osaka**
**My name is Taro Shimizu.**

All right.

We'll send you the announcement of the conference, so please refer to it.
Your name and address, please?

**Adam Smith.**
**My address is 2-27-7 Tamatsukuri, Higashi-ku, Osaka**

All right.
Wed like to ask your phone number also.

**Yes.**
**372-8018.**

372-8018, right?.

**Yes.**
**That's right.**
**Thank-you very much.**
**Good-bye.**

# Address Ellicitation in the Future

We'll send you the announcement of the conference, so please refer to it.
Your name and address, please?

**Oveissi Mohammed**

Oveissi Mohammed And is "Mohammed" your last name?

**No. Sorry. Oveissi is my last name.**

Would you please spell that?

**Yes. That's O-V-E-I**

O-V-E-I. umhmm.

**S-S-I. Oveissi.**

S-S-I. Oveissi
And the first name was Mohammed?.

**Yes. Mohammed.**

All right. And your mailing address.

B-R- **My address is Grona gatan 7**

**No. G-R-O-N-A. And gatan is G-A-T-A-N. And it's number 7**

ok. Grona gatan 7.

**And the city name is "Virserum" spelled V as in "Victor"-I-R**

V-I-R. ok.

**S-E-R-U-M as in "Mary"**

Allright. Virserum.

**The country is Sweden.**

And that's in Sweden is it? Ok. And is there a postal code?

**Yes. It's 57 (pause) 027/**

57 027. All right. And wed like to ask your phone number also.

**Yes. It's 372-8018.**

372-8018, right?.

**Yes.That's right. Thank-you very much. Good-bye.**

# DSIDS

*"Detailed Symbolic Information Dialogue Segments"*

**Need for accuracy increases dramatically**

**Degree of accuracy increases dramatically**

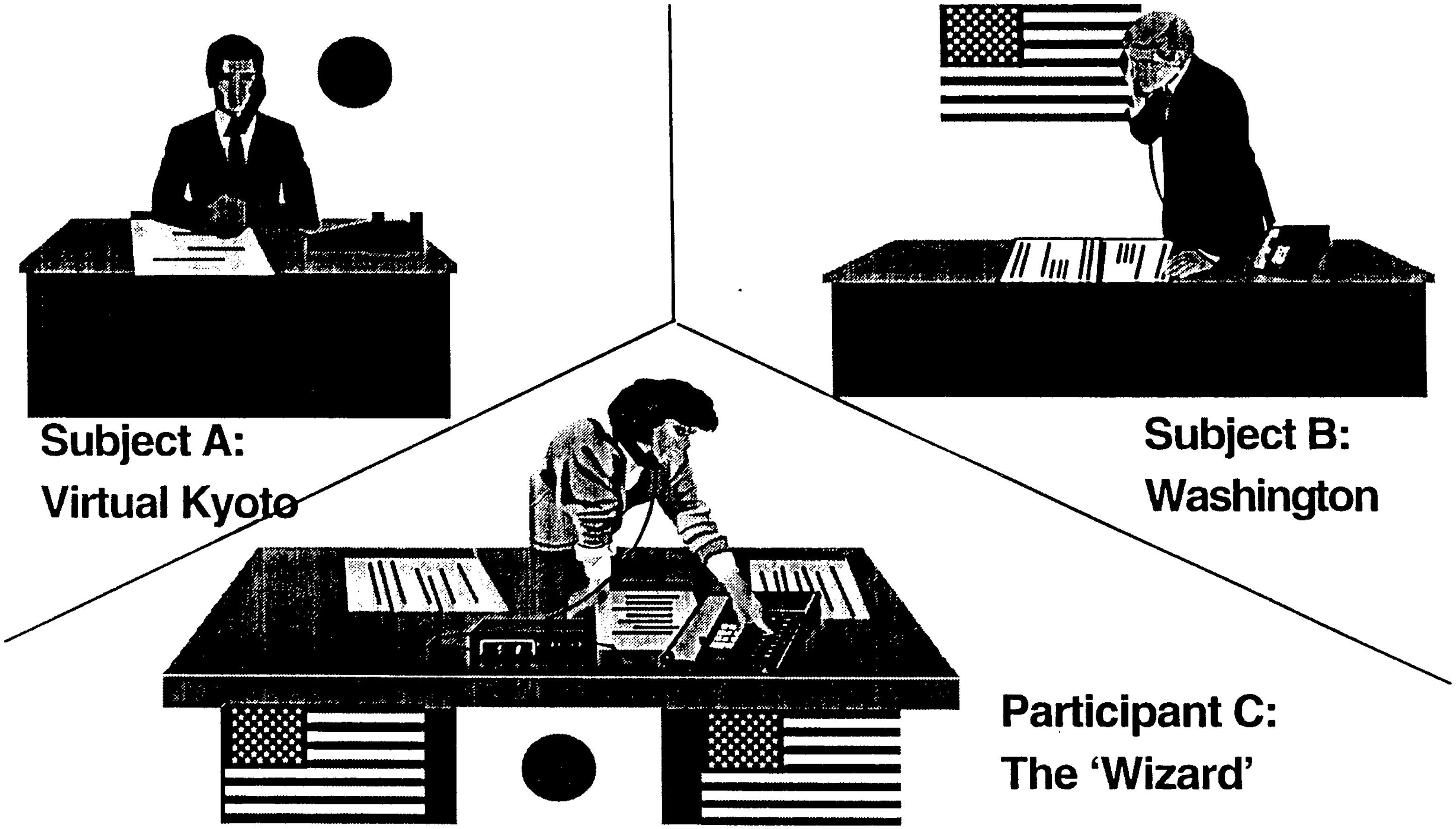
**Lexical expectations**
- **numerals**
- **letters**
- **help phrases "s as in Sam," 9--> "niner"**
- **domain-specific "Higashi," "Shi," "Road," "Department"**

**Packaging of information (Clark)**
- **pauses** *("five-seven <pause> o-five-seven")*
- **clarification subdialogues** *("Was that M as in Mary?")*
- **temporary suspension of MT** *("Department of Japan Studies")*

**Expect Proper Nouns and Novel terms**
*("Department of Japan Studies")*

**Opening and closing indicated at discourse level**

# Our Usability Study



Subject A:
Virtual Kyoto

Subject B:
Washington

Participant C:
The 'Wizard'

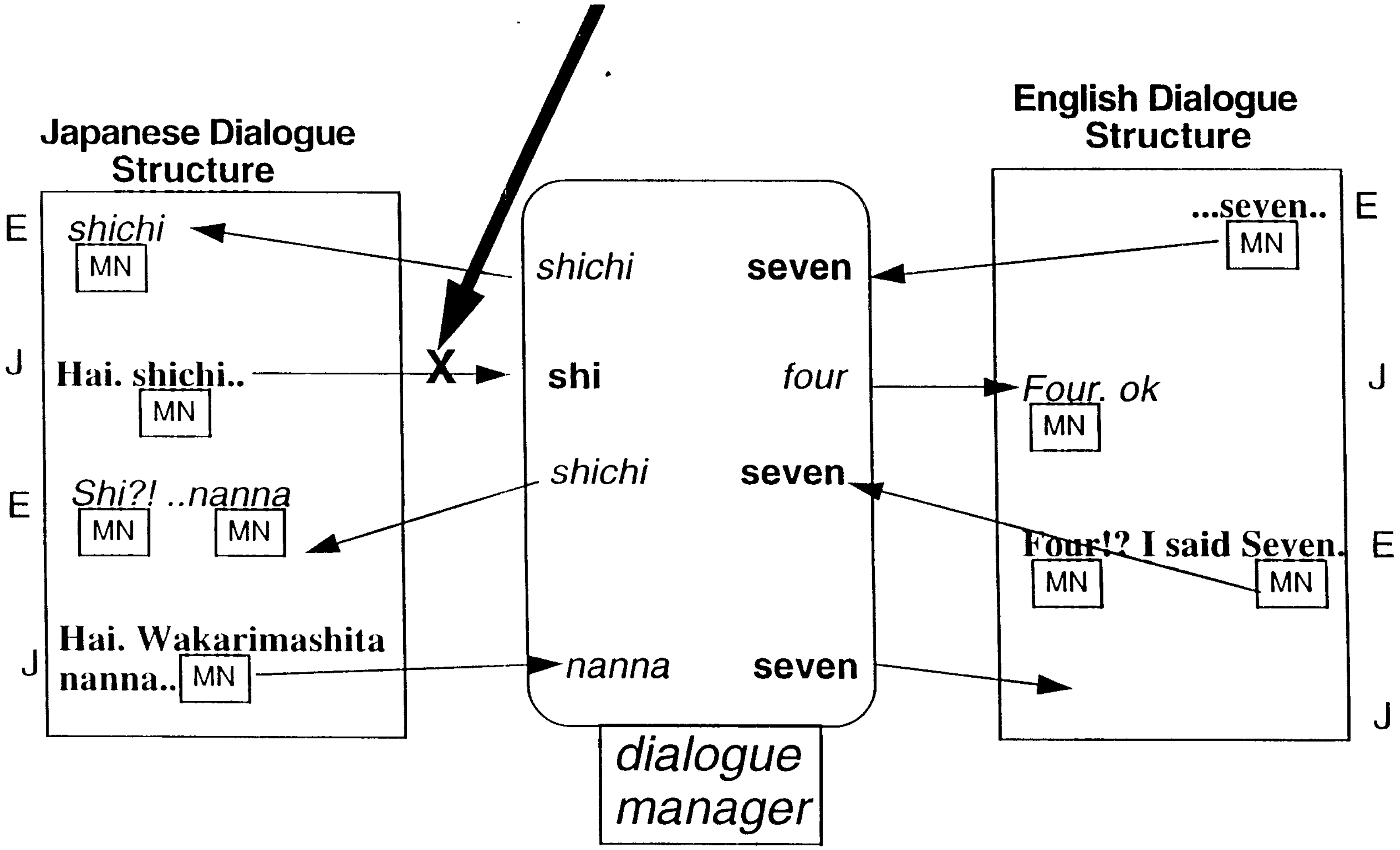
# Experimental Method

- Modified “Wizard of Oz” Design
  - electronic voice modulator
  - subjects sequestered
  - scripted repertoire of wizard behaviors
  - subjects contact wizard through “Phone” and “#”-key
- Execution
  - conference registration task
  - rehearsal with conventional telephone dialogue
  - subjects use native language at all times
  - video taped segments from both sides of dialogue
- Analysis
  - debriefing interview with each subject
  - transcription and informal coding of video taped data

# Empirical Study Results

- Speech-only interaction was inadequate
  - symbolic data and novel words
  - users wanted to know what was going on
  - users couldn't keep track of dialogue structure
- User-machine dialogue was essential
  - users need to address the system and vice versa
  - users make errors
  - backend system fails
  - users needed a way to say "done" after each input
- Half-duplex transmission was too slow, even with fastest possible intelligent agent

# Error Recovery



# Sample Discourse: Text Monologue

Operational testing is the final phase. At this point an independent group evaluates the system to verify its utility and production worthiness. There are, in fact, three simultaneous evaluations taking place in this phase. One is an evaluation of the equipment itself. The second is an evaluation of the early work done by the planners (who imagined what the utility would be if such a system were built). The third is an evaluation of the process that translated the vision of those operational planners into a specification. While we make three evaluations during operational testing, the last two could have been made years earlier if the system or some approximation of it had been available. In the sense that two of the three evaluations have experienced long and

# Sample Discourse: Spoken Dialogue

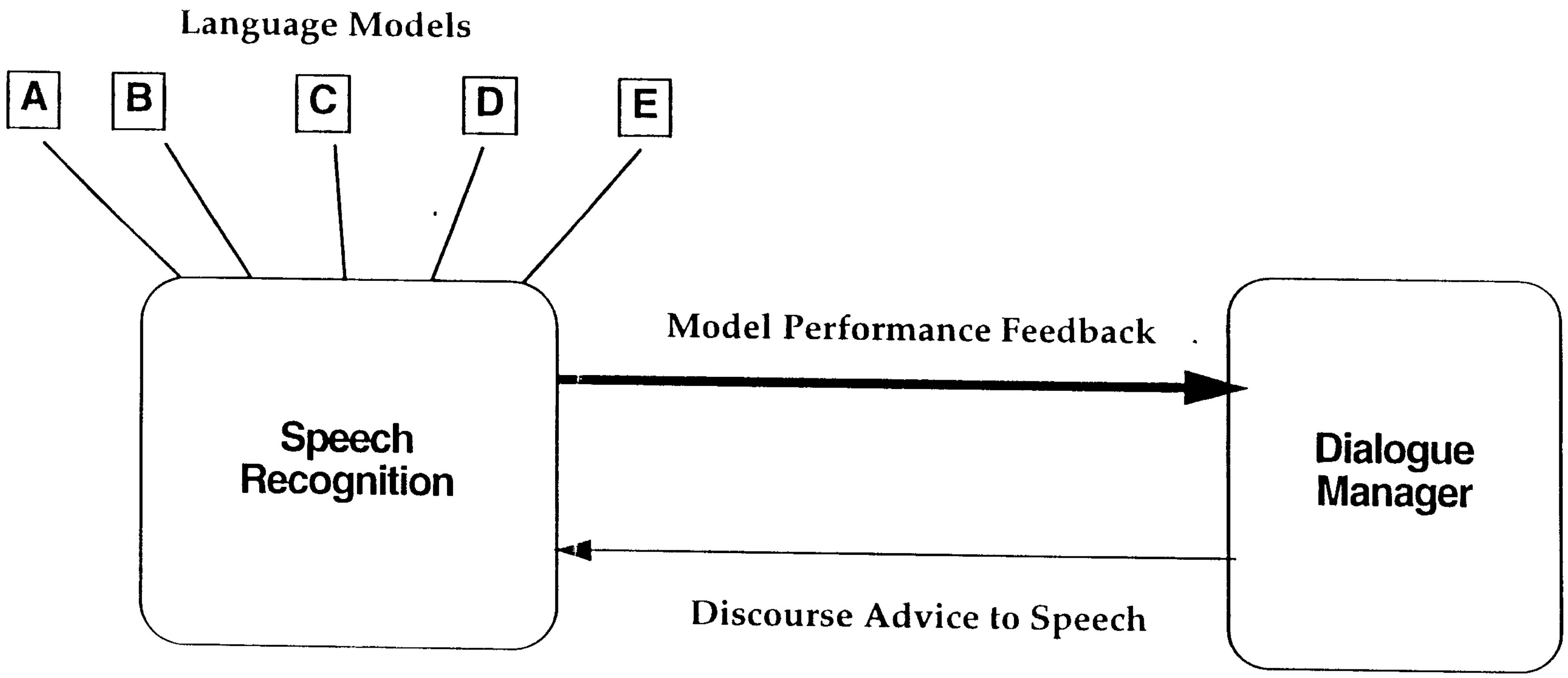
- I'd like to fly from Washington to Boston.
- Next Thursday
- Round trip
- Could you repeat that please?
- The following Sunday.

MITRE

# Fundamental Discourse Tasks

- Use stored discourse representation to improve interpretation of input and generation of output
  - anaphora resolution, ellipsis reconstruction, etc.
  - discourse planning, context-dependent form generation
- Update stored discourse representation
  - dialogue history
  - discourse state variables
- Provide discourse information to other modules

# Run-Time Protocol for Communication Between Discourse and Speech



Language Models

A

B

C

D

E

Speech Recognition

Model Performance Feedback

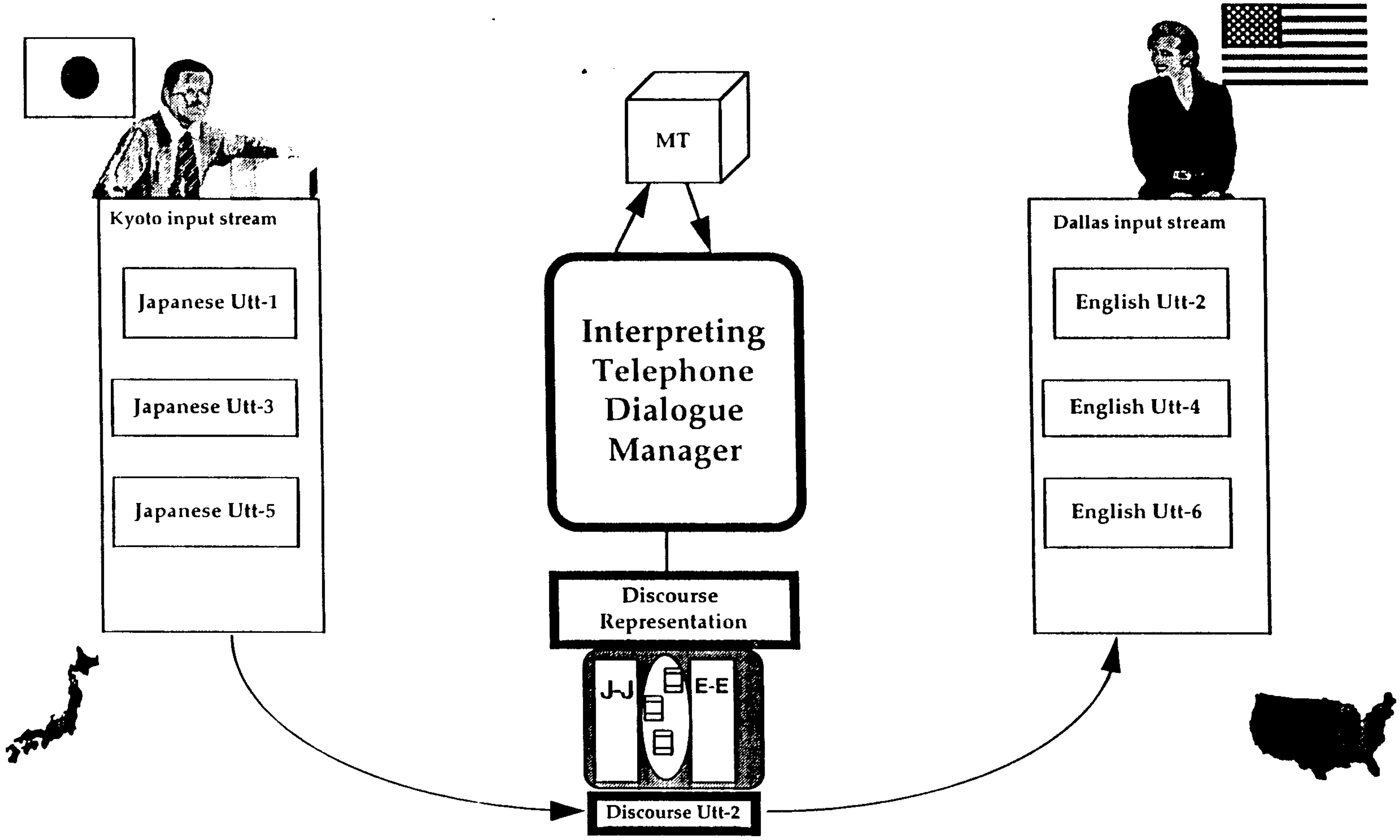
Discourse Advice to Speech

Dialogue Manager

# 4 Design Filters

1. The Data
   - linguistic phenomena that will occur
2. Upstream Modules
   - input providers
   - which information will be preserved
3. Downstream Modules
   - output consumers
   - which information can be made use of
4. Definition of Success
   - evaluation metrics

# Real-time Japanese-English Dialogue Translation



Kyoto input stream

Japanese Utt-1

Japanese Utt-3

Japanese Utt-5

MT

Interpreting Telephone Dialogue Manager

Discourse Representation

J-J

E-E

Discourse Utt-2

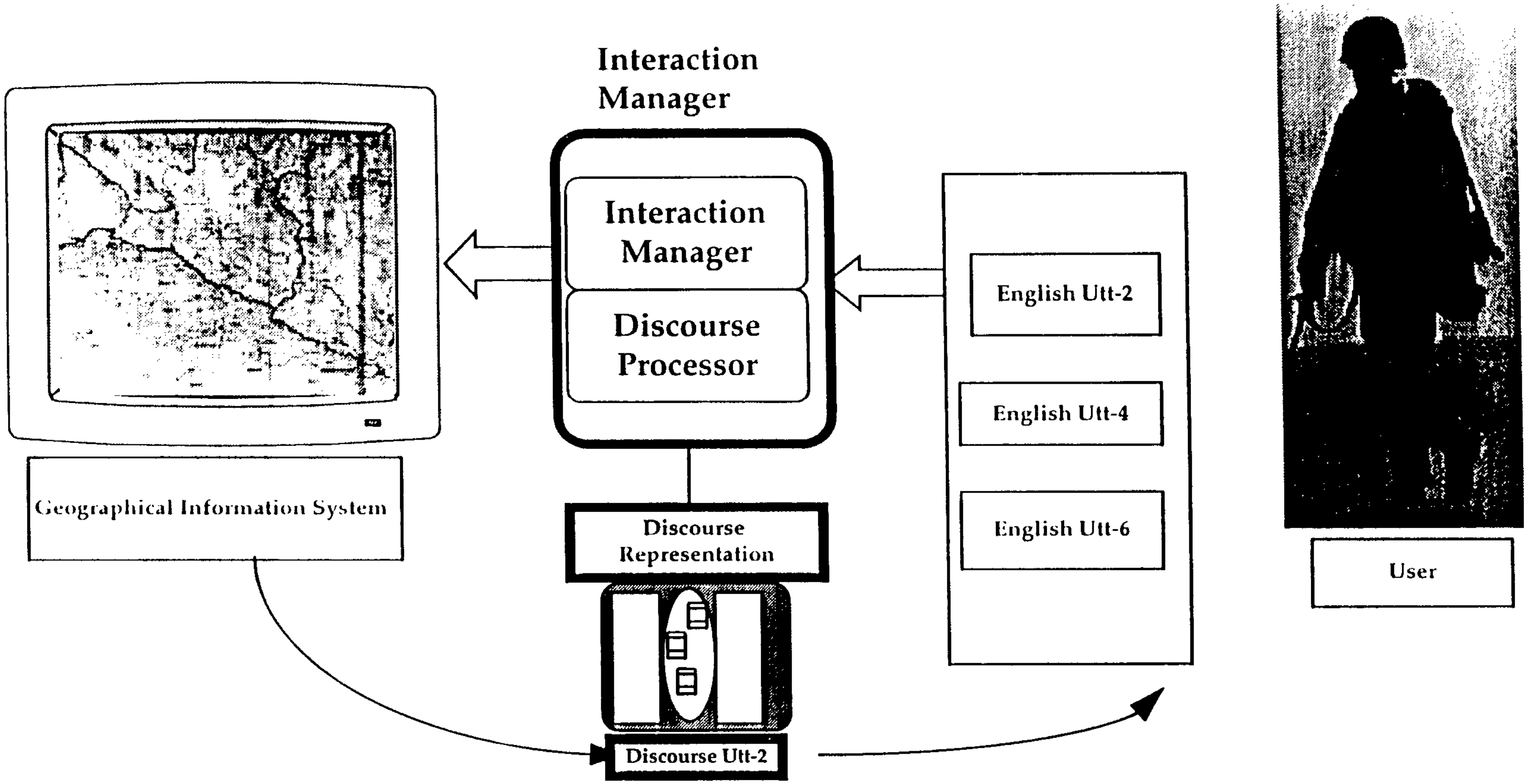
Dallas input stream

English Utt-2

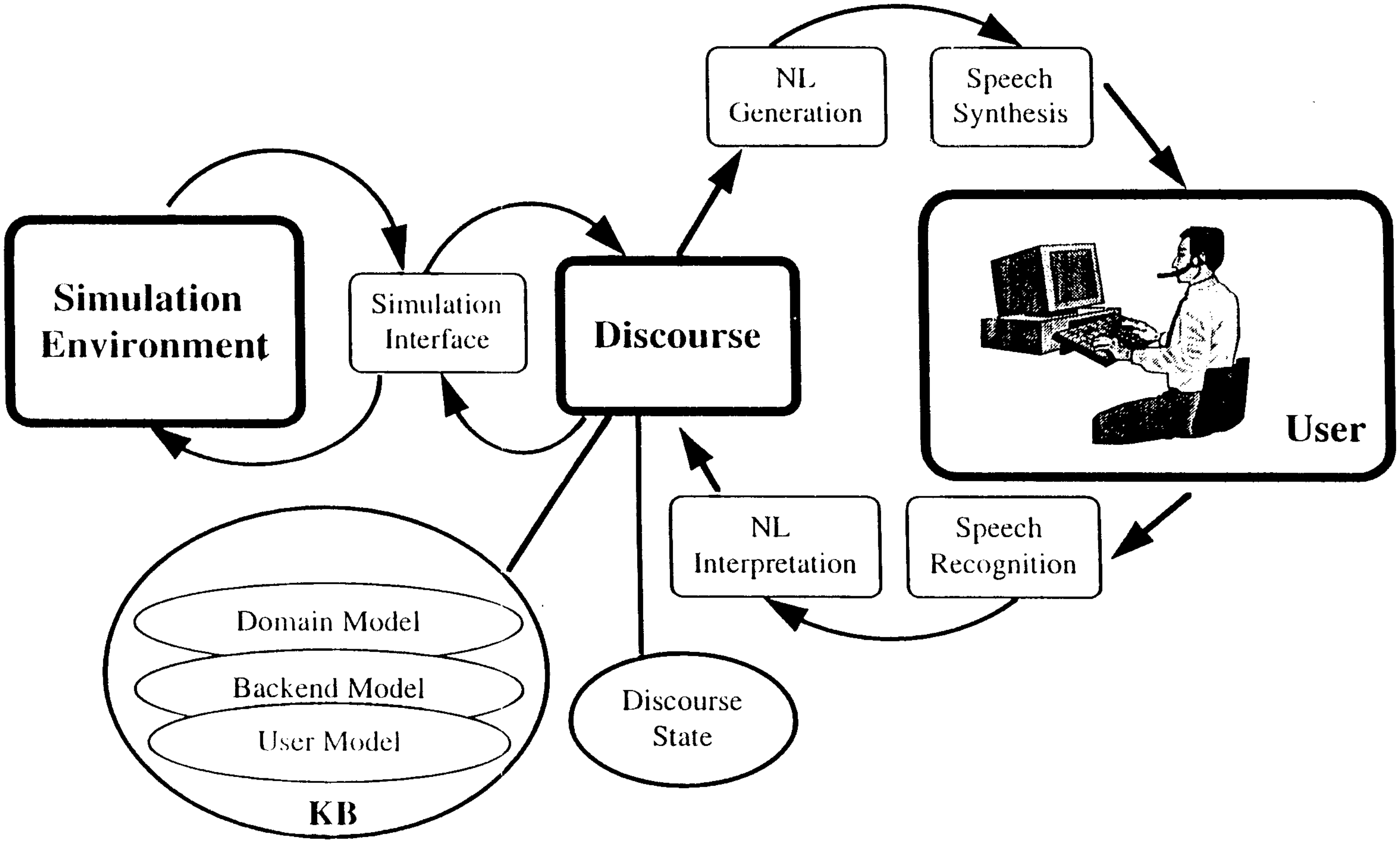
English Utt-4

English Utt-6

# MT View of Discourse Processing for Human-Computer Dialogue



# Distributed Simultation Interface



UNITED STATES SPECIAL OPERATIONS COMMAND

# Technology Review: Special Operations Forces (SOF) Speech Recognition for Language Sustainment

## Descriptions of Speech and Text Translation Systems

AIRBORNE

ARMY RESEARCH INSTITUTE

## JANUS: Spontaneous Speech to Speech Translation Environment Technology

Dr. Alex Waibel
Dr. Arthur E. McNair

The JANUS system will be demonstrated in two forms: as a translating videophone using workstations, and as a portable translation unit on a PC laptop. The demonstrated domain for translation is a scheduling task (communication between two humans to agree on a time/date to meet), though all technologies used are applicable to any domain, with effort currently required only to retrain the recognizer and build grammars for a new task or language (any overlap in tasks, such as dates, allows direct reuse of portions of grammars). The technologies demonstrated in JANUS include a spontaneous speech, speaker independent recognizer which can be trained for any language (currently English, German, Spanish, and Korean). Also used is a text-to-text translation system which uses hand-written grammars to parse input language text, and then generates text in multiple output languages (currently English, German, Spanish, Korean, and Japanese). Our current specialties include spontaneous speech recognition, multiple parsing/generation technologies (including automatic grammar generation), non-standard modes of human input to computers (speech, touch, handwriting, visual), and the combination of multiple input modalities in single applications.

Contact: Dr. Alex Waibel
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
(412) 268-7676; FAX: (412) 268-5578
e-mail: ahw@cs.cmu.edu

In Proceedings of the Theoretical and
Methodical Issues in Machine Translatio
Conf., Leuven, Belgium, July 5-7, 1995.

# Using Context in
# Machine Translation of Spoken Language

Lori Levin†, Oren Glickman†, Yan Qu†, Donna Gates†,
Alon Lavie†, Carolyn P. Rosé†, Carol Van Ess-Dykema‡, Alex Waibel†
† Carnegie Mellon University (USA)
‡ U.S. Department of Defense
lori.levin@nl.cs.cmu.edu

**Abstract:** We report on techniques for using discourse context to reduce ambiguity and improve translation accuracy in a multi-lingual (Spanish, German, and English) spoken language translation system. The techniques involve statistical models as well as knowledge-based models including discourse plan inference. This work is carried out in the context of the Janus project at Carnegie Mellon University and the University of Karlsruhe.

## 1 Introduction

Machine Translation of spoken language encounters all of the difficulties of written language (such as ambiguity) with the addition of problems that are specific to spoken language such as speech disfluencies, errors introduced during speech recognition, and the lack of clearly marked sentence boundaries. Fortunately, however, we can take advantage of the structure of task-oriented dialogs to help reduce these difficulties. In this paper we report on techniques for using discourse context to reduce ambiguity and improve translation accuracy in a multi-lingual (Spanish, German, and English) spoken language translation system. The techniques involve statistical models as well as knowledge-based models including discourse plan inference. This work is carried out in the context of the Janus project at Carnegie Mellon University and the University of Karlsruhe ([1]).

There has been much recent work on using context to constrain spoken language processing. Most of this work involves making predictions about possible sequences of utterances and using these predictions to limit the search space of the speech recognizer or some other component (See [2], [3], [4], [5], [6], [7], [8], [9]). The goal of such an approach is to increase the accuracy of the top best hypothesis of the speech recognizer, which is then passed on to the language processing components of the system. The underlying assumption being made is that design and complexity considerations require that each component of the system pass on a *single* hypothesis to the following stage, and that this can achieve sufficiently accurate translation results. However, this approach forces components to make disambiguation choices based solely on the level of knowledge available at that stage of processing. Thus, components of the system further down the line cannot correct a wrong choice of an earlier component.

The work reported in this paper does not rely on predictions about subsequent utterances (although we use such predictions in other work not reported here). The

| | |
|---|---|
| **s1:** qué te parece el lunes | *how do you feel about Monday?* |
| **s2:** tal vez sería mejor en la tarde<br>como a las a las dos de la tarde | *the afternoon is perhaps better*<br>*around two p.m.* |
| **s1:** no<br>yo tengo toda la tarde ocupada<br>de una a cuatro tengo una reunión | *no*<br>*i am busy all afternoon*<br>*from one o'clock till four o'clock i have a meeting* |
| **s2:** el lunes<br>entonces sería mejor el jueves | *Monday*<br>*then Thursday is better* |

Figure 1: **Example of Translation**

key feature of our approach is to allow multiple hypotheses to be processed through the system, and to use context to disambiguate between alternatives in the final stage of the process, where knowledge can be exploited to the fullest. Since it is infeasible to process *all* hypotheses produced by each of the system components, context is also used locally to prune out unlikely alternatives. We describe four approaches to disambiguation, two of which are sentence-based and two of which are discourse-based in that they take a multi-sentence context into account. We show that the use of discourse context improves performance on disambiguation tasks.

# 2 System Description

Janus is a speech-to-speech translation system currently dealing with dialogs in the scheduling domain (two people scheduling a meeting with each other). The current source languages are English, German, and Spanish and the target languages are English and German. We are also beginning to work with Korean, Japanese, and other languages. System development and testing is based on a collection of approximately 400 scheduling dialogs in each of the source languages. Translation of a portion of a transcribed dialog is shown in Figure 1.

The main modules of Janus are speech recognition, parsing, discourse processing, and generation. Each module is designed to be language-independent in the sense that it consists of a general processor that applies independently specified knowledge about different languages. Therefore, each module actually consists of a processor and a set of language-specific knowledge sources. A system diagram is shown in Figure 2.[1]

Processing starts with speech input in the source language. Recognition of the speech signal is done with acoustic modeling methods, constrained by a language model. The output of speech recognition is a word lattice. We prefer working with word lattices rather than the more common approach of processing N-best lists of hypotheses. An N-best list may be largely redundant and can be efficiently represented in the form of a lattice. Using a lattice parser can thus reduce time and space complexity relative to parsing a corresponding N-best list. Selection of the correct path through the lattice is accomplished during parsing when more information is available.

[1] Another approach being pursued in parallel in the Janus project is described in [10]

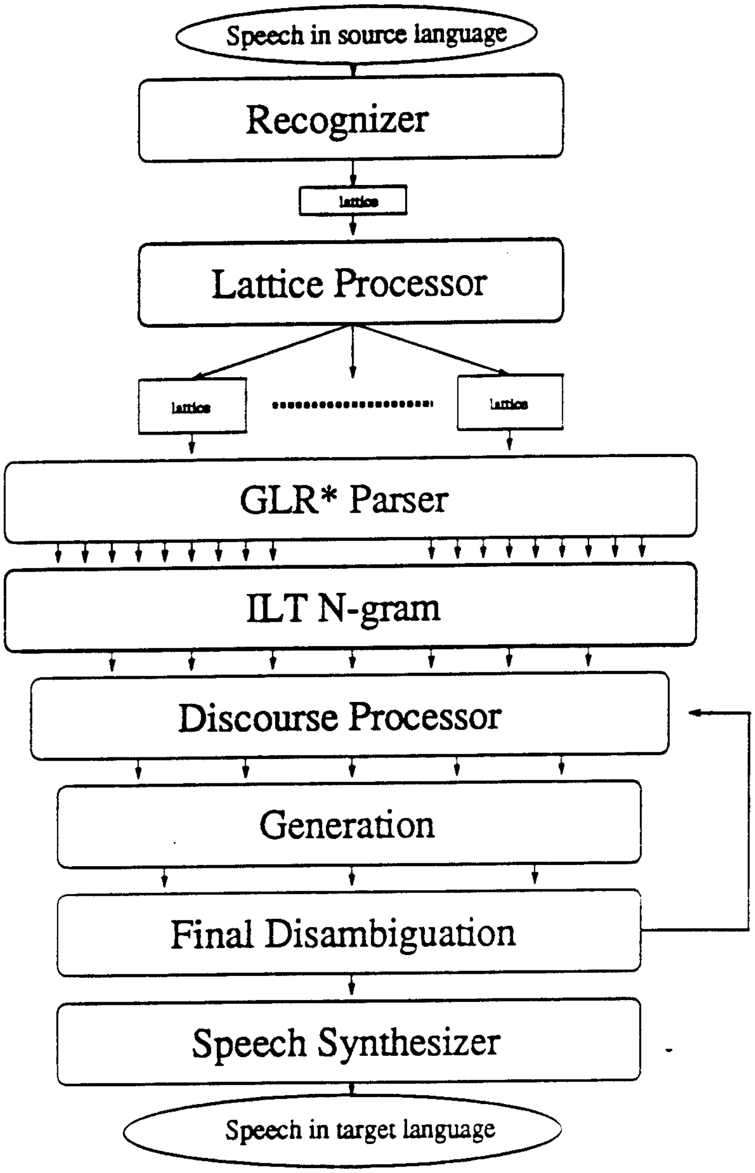


Figure 2: **Janus System Diagram**

Lattices, however, are potentially inefficient because of their size. We apply four steps to make them more tractable ([?]). The first step involves cleaning the lattice by mapping all non-human noises and pauses into a generic pause. Consecutive pauses are then adjoined to one long pause. The resulting lattice contains only linguistically meaningful information. The lattice is then broken at points where no human input is recognized over a specified threshold of time in the speech signal, yielding a set of sub-lattices which are highly correspondent to sentence breaks in the utterance. Each of the sub-lattices is then re-scored using a new language model. Finally the lattices are pruned to a size that the parser can process in reasonable time and space. The re-scoring raises the probability that the correct hypothesis will not be lost during the pruning stage. Each of the resulting sub-lattices are passed on to the parser, the first component of the translation process.

Parsing a word lattice involves finding all paths of connecting words within the lattice that are grammatical. The GLR* ([12], [13]) parser skips parts of the utterance that it cannot incorporate into a well-formed structure. Thus it is well-suited to domains in which extra-grammaticality is common. The parser can identify additional sentence breaks within each sub-lattice with the help of a statistical method that determines the probability of sentence breaks at each point in the utterance. The output of parsing a sub-lattice is a set of interlingua texts, or ILTs, representing all of the grammatical paths through the sub-lattice and all of the ambiguities in each grammatical path. The ILTs from each sub-lattice are combined, yielding a list of ILT sequences that represent the possible sentences of a full multi-sentence turn. An *ILT n-gram* is applied to each such list to determine the probability of each sequence of sentences.

The discourse processor, based on Lambert's work ([14, 15]), disambiguates the speech act of each sentence, normalizes temporal expressions, and incorporates the sentence into a discourse plan tree. The discourse processor's focusing heuristics and plan operators eliminate some ambiguity by filtering out hypotheses that do not fit into the current discourse context. The discourse component also updates a calendar in the dynamic discourse memory to keep track of what the speakers have said about their schedules.

As processing continues, the N-best hypotheses for sequences of ILTs in a multi-sentence turn are sent to the generator. The generation output for each of the N hypotheses is assigned a probability as well. The generation output follows certain forms and is restricted in style. Therefore a regular n-gram model can be applied to assign a probability to each hypothesis.

The final disambiguation combines all knowledge sources obtained: the acoustic score, the parse score, the ILT n-gram score, information from the discourse processor, and a generation n-gram score. The best scoring hypothesis is sent to the speech synthesizer. This hypothesis is also sent back to the discourse processor so it can update its internal structures and the discourse state accordingly.

During translation, several knowledge structures are produced which constitute a discourse context that other processes can refer to. These knowledge structures include the ILT, the plan tree and focus stack, and the dynamically produced calendar. The main components of an ILT are the speech act (e.g., `suggest`, `accept`, `reject`), the sentence type (e.g., `state`, `query-if`, `fragment`), and the main semantic frame

"Estás ocupada el lunes"
*(Are you busy on Monday)*

```
((FRAME *BUSY)
 (SENTENCE-TYPE *QUERY-IF)
 (A-SPEECH-ACT (*MULTIPLE* *SUGGEST
                           *REQUEST-RESPONSE))
 (SPEECH-ACT *REQUEST-RESPONSE)
 (WHO ((FRAME *YOU)))
 (WHEN
      ((WH -) (FRAME *SIMPLE-TIME)
              (SPECIFIER DEFINITE)
              (DAY-OF-WEEK MONDAY))))
```

Figure 3: An Interlingua Text (ILT)

(e.g., `free`, `busy`). An example oƒ an ILT is shown in Figure 3. The plan tree is based on a three-level model of discourse with discourse, domain, and problem solving levels. It shows how the sentences relate to each other in discourse segments. The focus stack indicates which nodes in the plan tree are available for further attachments. Figure 4 shows a plan tree at the discourse level. The first sentence, which is a surface question, is identified as a `Ref-Request` (request for information), a `Suggest-Form` (a possible way of making a suggestion), and finally part of an `Obtain-Agreement-Attempt` (a portion of the discourse in which the two speakers attempt to come to some agreement). The next sentence attaches as a `Self-Initiated-Clarification` indicating that this sentence makes the suggestion in the previous sentence more clear. The last two sentences are both `Accept-Forms` (acceptance of a suggestion) which chain up together to a `Response` node which then attaches to the corresponding suggestion. The Calendar records times which the speakers are considering, suggesting, rejecting, etc. This is updated dynamically as the conversation progresses. An example of a calendar is shown in Figure 5. Procedures that resolve ambiguity and select from among alternative analysis can take advantage of these knowledge structures as well as simpler ones such as the words in the previous sentence.

# 3 Techniques for Disambiguation

Resolution of ambiguity is important for accurate translation. Table 1 shows some examples of translation errors that are caused by failure to resolve ambiguity correctly. This section describes four disambiguation methods differing along two dimensions, whether they are knowledge-based or statistical, and whether they are sentence-based or take discourse context into account. The different types of ambiguities encountered in Spanish-to-English translation are summarized in Figure 6.

The following subsections describe the disambiguation methods that we tested. Our sentence-based disambiguation methods are implemented within the GLR* parser ([12] [13]) and its accompanying grammar. One method is knowledge-based, involving preferences that are explicitly encoded in grammar rules. The other is statistical, involving probabilities of actions in the LR parsing table. The context-based methods

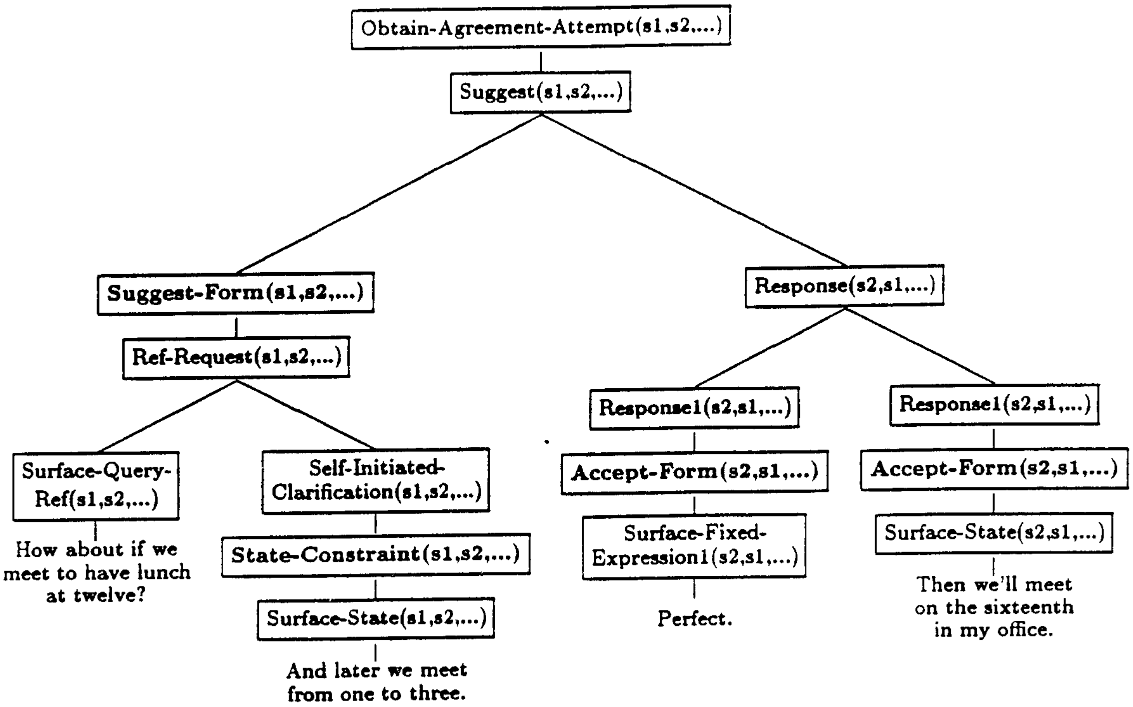
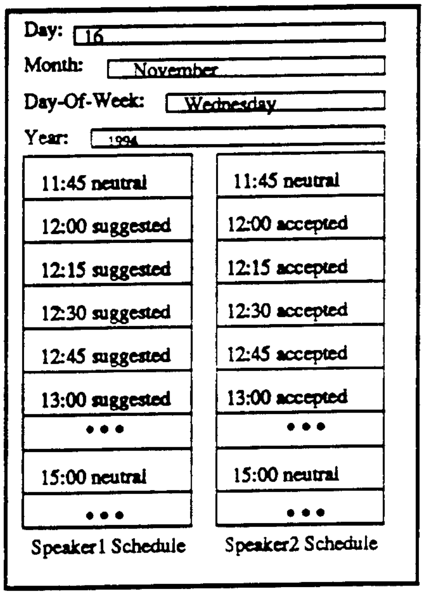

Obtain-Agreement-Attempt(s1,s2,...)
Suggest(s1,s2,...)
Suggest-Form(s1,s2,...)
Response(s2,s1,...)
Ref-Request(s1,s2,...)
Response1(s2,s1,...)
Response1(s2,s1,...)
Surface-Query-Ref(s1,s2,...)
Self-Initiated-Clarification(s1,s2,...)
Accept-Form(s2,s1,...)
Accept-Form(s2,s1,...)
How about if we meet to have lunch at twelve?
State-Constraint(s1,s2,...)
Surface-Fixed-Expression1(s2,s1,...)
Surface-State(s2,s1,...)
Surface-State(s1,s2,...)
Perfect.
Then we'll meet on the sixteenth in my office.
And later we meet from one to three.

Figure 4: Example Plan Tree



Day: 16

Month: November

Day-Of-Week: Wednesday

Year: 1994

| | |
|---|---|
| 11:45 neutral | 11:45 neutral |
| 12:00 suggested | 12:00 accepted |
| 12:15 suggested | 12:15 accepted |
| 12:30 suggested | 12:30 accepted |
| 12:45 suggested | 12:45 accepted |
| 13:00 suggested | 13:00 accepted |
| ••• | ••• |
| 15:00 neutral | 15:00 neutral |
| ••• | ••• |
| Speaker1 Schedule | Speaker2 Schedule |

Figure 5: A Calendar Day Structure

| Spanish Input | Actual Translation | Correct Translation |
|---|---|---|
| Example 1<br>**s1:** hola Patricia<br>cómo estás | hello – Patricia<br>How do you feel about it? | *How are you?* |
| Example 2<br>**s1:** en la tarde del miércoles<br>**s2:** bueno<br>dame un poquito de tiempo para reunirme contigo<br>**s1:** qué tal de dos a cuatro<br>**s2:** fabuloso | Wednesday afternoon<br>okay<br>give me a little time to meet with you<br>how about from the second till the fourth?<br>that sounds great | *how about from two o'clock till four o'clock?* |
| Example 3<br>**s1:** así que si tiene alguna hora en esos días será mejor | so if you are free at some time – those days are better | *so if you are free at some time on those days – that is better* |

Table 1: **Mistranslations of Ambiguous Sentences**

include knowledge-based discourse plan inference and statistical N-grams of ILTs.

## Parse Disambiguation Using Grammar Rule Preferences

In order to successfully parse fragmented input, the grammars we use for parsing spontaneous speech have very inclusive notions as to what may constitute a "grammatical" sentence. The grammars allow meaningful clauses and fragments to propagate up to the top (sentence) level of the grammar, so that fragments may be considered complete sentences. Additional grammar rules allow an utterance to be analyzed as a collection of several grammatical fragments. The major negative consequence of this grammar "looseness" is a significant increase in the degree of ambiguity of the grammar. In particular, utterances that can be analyzed as a single grammatical sentence, can often also be analyzed in various ways as collections of clauses and fragments. Our experiments have indicated that, in most such cases, a less fragmented analysis is more desirable. Thus, we developed a mechanism for prefering less fragmented analysis.

The fragmentation of an analysis is reflected via grammar preferences that are set explicitly in various grammar rules. The preferences are recorded in a special *counter* slot in the constructed feature structure. By assigning counter slot values to the interlingua structure produced by rules of the grammar, the grammar writer can explicitly express the expected measure of fragmentation that is associated with a particular grammar rule. For example, rules that combine fragments in less structured ways can be associated with higher counter values. As a result, analyses that are constructed using such rules will have higher counter values than those constructed with more structurally "grammatical" rules, reflecting the fact that they are more fragmented. Although used to primarily reflect preferences with respect to fragmentation, the same mechanism can be used to express other preferences as well.

We tested the disambiguation performance of the GLR* parser using the grammar preferences as the sole disambiguation criterion. In this setting, for an ambiguous sentence that results in multiple analysis, the parser chooses the analysis with the lowest counter value. Ties between numerous analyses with equal minimal counter score are broken at random. This disambiguation method was tested on a set of 512 sentences, 252 of which produce ambiguous parses. As shown in Table 2, the GLR* parser selected the correct parse in 196 out of the 252 ambiguous sentences. This corresponds to a success rate of 78%.

## Parse Disambiguation Using a Statistical Model

The grammar rule preference mechanism can reflect preferences between particular grammar rules. However, it does not provide a complete mechanism for disambiguating between the set of all possible analyses of a given input. This is done by a statistical module which augments the parser. Our statistical model attaches probabilities directly to the alternative actions of each state in the parsing table. Because the state of the GLR* parser partially reflects the left and right context within the sentence of the parse being constructed, modeling the probabilities at this level has the potential of capturing preferences that cannot be captured by standard Probabilistic Context-Free Grammars. For example, a reduce action by a certain grammar rule $A \rightarrow \alpha$ that appears in more than one state can be assigned a different probability in each of the occurrences.

Training of the probabilities is performed on a set of disambiguated parses. The probabilities of the parse actions induce statistical scores on alternative parse trees, which are then used for parse disambiguation.

We tested the disambiguation performance of the GLR* parser using a combination of the statistical parse scores and the grammar rule preference values. The same test set of 252 ambiguous sentences was evaluated. As can be seen in Table 2, the combined disambiguation method succeeds in selecting the correct parse in 209 of the 252 cases, a success rate of 82%.

## Disambiguation Using Discourse Plans

Our discourse processor is a plan inference model based on the recent work of Lambert ([14, 15]). The system takes as its input ILTs of sentences as they are uttered and relates them to the existing context, i.e., the plan tree. Plan inferencing starts from the surface forms of sentences. Then speech-acts are inferred. Multiple speech-acts for one ILT could be inferred. A separate inference chain is created for each possible speech act. Preferences for picking one inference chain over another are determined by the focusing heuristics, which provide ordered expectations of discourse actions given the existing plan tree. A detailed description of the focusing heuristics can be found in [16] and [17].

We are currently conducting experiments to see how the plan tree and focusing heuristics can help to disambiguate multiple ILT outputs from the parser. We have obtained some preliminary results concerning resolving ambiguities in sentence types (`statement`, `query-if`, `query-ref`, `fixed-expression`, `fragment`) in the ILT outputs. Our experiments have shown that the same focusing heuristics, which are useful for picking the most prefered inference chain for one ILT, can be used for providing

| Type of Ambiguity | Number of Occurences | Examples |
|---|---|---|
| **Slot**<br>A piece of information occurs in different slots in each ILT. | 20 | si estás libre el martes ocho puedo reunirme todo el día<br>*If you are free on Tuesday the eighth, I can meet all day.* or<br>*If you are free, on Tuesday the eighth I can meet all day.* or<br>*If you are free on Tuesday, on the eighth I can meet all day.*<br>voy a estar afuera la semana que viene<br>*I will be out of town the week that's coming up.* or<br>*I will be out of town the week that you're coming.*<br>este día<br>*this day* or *um day* |
| **Value**<br>The ILTs differ in the value of a slot. | 162 | nos podemos reunir a las dos<br>*We can meet at two.* or *Can we meet at two?*<br>nos reunimos el veintitrés<br>*We will meet on the twenty third.* or<br>*We met on the twenty third.*<br>dos a cuatro<br>*second at four* or *second to forth* or *two to four* |
| **Frame**<br>The ILTs have different top-level frames. | 136 | vamos a ver<br>*Let's see.* or *We will check.* or *We will see.*<br>bueno<br>*Good* or *Well...*<br>qué tal<br>*How are you?* or *How is that?* |
| **Sentence breaking**<br>The grammar allows more than one way of breaking the input into sentences. | 46 | el dos es bueno<br>*The second is good.* or *It is the second. Good.*<br>no está bien<br>*It is not good.* or *No, it is good.*<br>qué bueno<br>*How great!* or *What? Good.* |
| **Duplicate**<br>The parser produces multiple identical ILTs. | 31 | voy a salir a las dos probablemente<br>*I will leave on the second probably.*<br>el martes es el dos de octubre<br>*Tuesday is the second of October.* |
| **All types** | 395 | |

Figure 6: Types of Ambiguities

ordered expectations for picking inference chains from multiple ILT outputs of the parser.

The design of the experiment is composed of two steps. First, we try to attach each ILT from the set of ambiguous ILTs of a sentence to the existing dialog model. Second, the results of attachment for each ILT are compared. The best attachment is considered to be the one which best continues the existing context. When multiple attachments are possible, the focusing heuristics are used to make comparisons. For example, the sentence *Y nos podríamos reunir a la una* can be a statement (*And we could meet at one*) or yes-no question (*And could we meet at one?*). The focusing heuristic prefers the statement because it attaches to the current focus action, whereas the question attaches to an ancestor of the current focus action. The performance result of using plan tree and focusing strategy on sentence type ambiguities is shown in Table 3.

From Table 3, it can be seen that by using context and the focusing heuristics, the discourse processor achieves a general performance of 86% for sentence type disambiguation, which is an improvement over the 80% performance of the statistical parser without using context. For the `statement` vs `query-if` ambiguity, the discourse processor has a performance of 85%.

## Statistical Methods for Using Context for Disambiguation

As we described above, the statistical scores assigned by the parser are based on sentence structure without taking the context of surrounding sentences into account. In this section we describe a statistical approach that uses context to help parse disambiguation. This work involved assigning probabilities to full utterances. We consider a full utterance, U, as a sequence of sentences represented by ILTs. Such an utterance could be assigned an approximated bigram probability by the formula:

$$\Pr(U) = \Pr(\mathrm{ILT}_1, \mathrm{ILT}_2, \ldots, \mathrm{ILT}_n) = \prod_{i=1}^{n} \Pr(\mathrm{ILT}_i \mid \mathrm{ILT}_{i-1}) \qquad (1)$$

If $\mathrm{ILT}_i$ is the first ILT of an utterance, then $\mathrm{ILT}_{i-1}$ is the last ILT in the previous utterance of the other speaker.

Because we can not compute bigrams of full ILTs, our preliminary work has involved computing the probabilities of the `sentence-type`, `speech-act` and top-level `frame` of an ILT using the bigram probabilities described below. Standard smoothing techniques are used to calculate the conditional probabilities. Because we take into account the speakers of the current and previous sentences, a slot from the previous ILT is considered differently depending on if it was uttered by the same speaker or not. The amount of training data was not sufficient to calculate more complex N-grams such as $\Pr(\text{frame}_n \mid \text{frame}_{n-1}\ \text{sentence-type}_{n-1}\ \text{speech-act}_{n-1})$ or $\Pr(\text{frame}_n \mid \text{frame}_{n-1}\ \text{frame}_{n-2})$ . We thus compute only the following probabilities:

$P_1 = \Pr(\text{sentence-type}_n \mid \text{sentence-type}_{n-1})$
$P_2 = \Pr(\text{sentence-type}_n \mid \text{speech-act}_{n-1})$
$P_3 = \Pr(\text{sentence-type}_n \mid \text{frame}_{n-1})$

|  | Random | Grammar Preferences | Statistical Parse Disambiguation | ILT N-gram | Number of Sentences |
|---|---|---|---|---|---|
| **Cross-talk** | 41% | 81% | 84% | 88% | 91 |
| **Push-to-talk** | 39% | 76% | 81% | 83% | 161 |
| **Total** | 40% | 78% | 82% | 85% | 252 |

Table 2: **Disambiguation of All Ambiguous Sentences**

$P_4 = \Pr(\text{frame}_n \mid \text{sentence-type}_{n-1})$
$P_5 = \Pr(\text{frame}_n \mid \text{speech-act}_{n-1})$
$P_6 = \Pr(\text{frame}_n \mid \text{frame}_{n-1})$

The above probabilities together with the parser's score, $P_0$, are interpolated to assign the ILT's conditional probability $\Pr(\text{ILT}_n \mid \text{ILT}_{n-1}) = \sum_{i=0}^{6} \lambda_i P_i$, where the weights sum to one and are assigned so as to maximize the performance of the model.

# 4 Comparison of Disambiguation Methods

Each of the disambiguation methods described above was trained or developed on a set of thirty Spanish scheduling dialogs and tested on a set of fifteen previously unseen dialogs. The development set and test set both contain a mixture of dialogs that were recorded in two different modes. In push-to-talk dialogs, participants cannot interrupt each other. The speaker must hit a key to indicate that he or she is finished speaking before the other participant can speak. In cross-talk dialogs, the participants can interrupt each other and speak simultaneously. Each speaker is recorded on a separate track. Push-to-talk sentences tend to be longer and more complex.

Table 2 shows the performance of three disambiguation methods in comparison to a baseline method of selecting a parse randomly. The three disambiguation methods are cumulative in the sense that each one builds on the previous one. The first method, Grammar Preferences, involves the explicit coding of preferences in grammar rules. The second method, Statistical Parse Disambiguation, refers to the parse score computed by the GLR* parser, which takes into account the probabilities of actions in the GLR* parsing table as well as the grammar preferences. The third method, ILT n-grams, disambiguates top-level frames, sentence-types, and speech-acts, but relies on the parse score to resolve other ambiguities. As can be seen in Table 2 and Figure 7, each method adds a slight improvement over the others that it incorporates.

Table 3 shows the performance of four disambiguation methods in resolving sentence-type ambiguities. The first row shows performance on the most common ambiguity in Spanish—the ambiguity between statements and yes-no questions (`query-if`). Without access to intonation, statements are often indistinguishable from yes-no questions because they have the same word order in some circumstances. The four methods compared are the Grammar Preferences, Statistical Parse Disambiguation, and ILT N-grams described above, as well as Discourse Plan Inference. The Discourse Plan Inference is not cumulative with the other disambiguation methods. The input to the

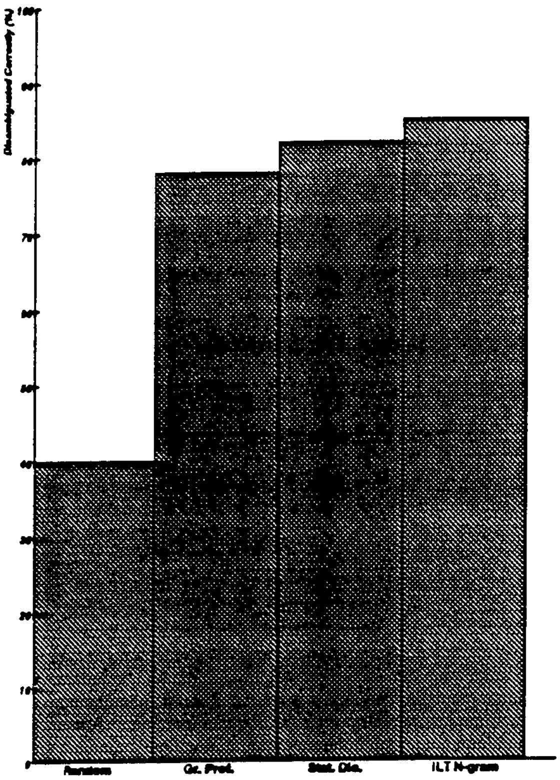


Figure 7: **Disambiguation of All Ambiguous Sentences**

| | Random | Grammar Preferences | Statistical Parse Disambiguation | Discourse Plans | ILT N-gram | Number of Sentences |
|---|---|---|---|---|---|---|
| **Statement/Query-if Ambiguity** | 57% | 82% | 80% | 85% | 94% | 114 |
| **All Sentence Type Ambiguities** | 51% | 82% | 80% | 86% | 90% | 166 |

Table 3: **Disambiguation of Sentence Types**

plan inference system is all of the ambiguous ILTs from the parser, without statistical parse scores. In this table, performance is calculated for the correct disambiguation of sentence-type only. Other ambiguities in the same sentences are not counted. The context-based methods, ILT N-grams and Discourse Plan Inference, perform better than the sentence-based methods in resolving the ambiguity between statements and yes-no questions. The second row of the table shows performance on all sentence-type ambiguities. Here also, the context-based methods do better than the sentence-based methods.

# 5 Conclusion

The approach we have taken is to allow multiple hypotheses and their corresponding ambiguities to cascade through the translation components, accumulating information that is relevant to disambiguation along the way. In contrast to other approaches that use predictions to filter out ambiguities early on, we delay ambiguity resolution as much as possible until the stage at which all knowledge sources can be exploited. A consequence of this approach is that much of our research effort is devoted to the development of an integrated set of disambiguation methods that make use of statistical and symbolic knowledge.

In this paper we examined four disambiguation methods, two that are sentence-based and two that use discourse context. In our experiments, the context-based methods performed somewhat better than the sentence-based methods. However, we believe that the best approach will be an integration of these and possibly other methods. Our future work will involve in particular how to combine the knowledge provided by the discourse processor with that provided by the parser and ILT N-grams. We believe that this is a promising path to follow because different sets of sentences are correctly disambiguated by each of the methods. Another feature of our future work will be to evaluate the effect of improved disambiguation on overall end-to-end translation quality.

# References

[1] M. Woszczyna, N. Aoki-Waibel, F.D. Buo, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel. JANUS 93: Towards Spontaneous Speech Translation, In *ICASSP*, 1994.

[2] S. R. Young. Use of Dialog, Pragmatics and Semantics to Enhance Speech Recognition, *Speech Communication*, 9, 1990, pages 551-564.

[3] W. Ward, S. Young. Flexible Use of Semantic Constraints in Speech Recognition, In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, 1993, Vol. 2, pp. 49–50

[4] S. R. Young, A. G. Hauptmann, W. H. Ward, E. T. Smith, P. Werner. High Level Knowledge Sources in Usable Speech Recognition Systems, *Communications of the ACM*, Volume 32 Number 2, February 1989, p 183 - 194.

[5] H. Iida, T. Yamaoka, H. Arita. Predicting the Next Utterance Linguistic Expressions Using Contextual Information, *IEICE Trans. Inf. & Suyst.*, Vol. E76-D, No. 1, January 1993.

[6] N. Reithinger, E. Maier. Utilizing Statistical Dialogue Act Processing in Verbmobil In *Proceedings of the ACL*, 1995.

[7] N. Reithinger. Some Experiments in Speech Act Prediction In *Proceedings of AAAI Spring Symposium*, 1995.

[8] J. Alexandersson, E. Maiser, N. Reithinger. A Robust and Efficient Three-Layered Dialogue Component for a Speech-to-Speech Translation System In *Proceedings of the EACL*, Dublin, 1995.

[9] G. Hanrieder, G. Görz. Robust Parsing of Spoken Dialogue Using Contextual Knowledge and Recognition Probabilities To appear in *Proceedings of ESCA Workshop on Spoken Dialogue Systems*, 1995.

[10] L.J. Mayfield, M. Gavalda, Y-H. Seo, B. Suhm, W. Ward, A. Waibel Concept-Based Parsing For Speech Translation, In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, 1995.

[11] Oren Glickman. Using Domain Knowledge to Improve End to End Performance in a Speech Translation System, Technical Report, Laboratory for Computational Linguisitcs, Carnegie Mellon University, 1995.

[12] A. Lavie, M. Tomita. GLR* - An Efficient Noise-skipping Parsing Algorithm for Context-free Grammars, In Proceedings of Third International Workshop on Parsing Technologies, 1993, pp. 123–134

[13] A. Lavie. An Integrated Heuristic for Partial Parse Evaluation, In *Proceedings of 32nd Annual Meeting of the ACL*, 1994

[14] L. Lambert and S. Carberry. Modeling negotiation subdialogues. In *Proceedings of the ACL*, 1992.

[15] L. Lambert. *Recognizing Complex Discourse Acts: A Tripartite Plan-Based Models of Dialogue.* PhD thesis, University of Delaware, September 1993.

[16] C. P. Rosé. Plan-based discourse processor for negotiation dialogues. Unpublished ms, January 1995.

[17] C. P. Rosé, B. Di Eugenio, L. S. Levin, C. Van Ess-Dykema. *Discourse Processing of Dialogues with Multiple Threads* To appear in *Proceedings of the ACL*, 1995.

In Proceedings of the ARPA Spoken Language Technology Workshop, Austin, TX, January 1995.

# JANUS: TOWARDS MULTILINGUAL SPOKEN LANGUAGE TRANSLATION

*B. Suhm*[1], *P. Geutner*[2], *T. Kemp*[2], *A. Lavie*[1], *L. Mayfield*[1], *A. E. McNair*[1], *I. Rogina*[2], *T. Schultz*[2], *T. Sloboda*[2], *W. Ward*[1], *M. Woszczyna*[1], *A. Waibel* [1,2]

Interactive Systems Laboratories
[1] Carnegie Mellon University (USA)
[2] Karlsruhe University (Germany)

## ABSTRACT

In our effort to build spoken language translation systems we have extended our JANUS system to process spontaneous human–human dialogs in a new domain, two people trying to schedule a meeting. Trained on an initial database JANUS-2 is able to translate English and German spoken input in either English, German, Spanish, Japanese or Korean output. To tackle the difficulty of spontaneous human–human dialogs we improved the JANUS-2 recognizer along its three knowledge sources acoustic models, dictionary and language models. We developed a robust translation system which performs semantic rather than syntactic analysis and thus is particulary suited to processing spontaneous speech. We describe repair methods to recover from recognition errors.

## 1. Introduction

JANUS [1, 2] has been among the first systems attempting to provide spoken language translation. While the previous JANUS-1 system processed syntactically wellformed read speech over a 500 word vocabulary, JANUS-2 operates on spontaneous human–human dialogs in a scheduling domain with vocabularies exceeding 2000 words. Currently, English and German spoken input can be translated in either English, German, Spanish, Japanese or Korean output. Work is in progress to add Spanish and Korean as input languages.

This paper reports on the current status of the system and ongoing efforts to extend and improve the recognition component. Then, we describe our new approach to robust translation of spoken language. We briefly describe and compare the alternative approach to parsing and translation we pursue, based on a generalized robust LR parser and an ILT. Finally we report on efforts to detect erroneous system output and provide interactive methods to recover from such errors.

## 2. Current Status of JANUS

### 2.1. Data Collection

Data collection to establish a large database of spontaneous human–human negotiation dialogs in English and German has started about 18 months ago. In the meantime, several sites in Europe, the US and Asia have adopted the Scheduling task under several research projects and funding sources. Since the same calendars and data collection protocols are used the data elicited shares the same domain and procedural constraints.

| English Scheduling | | |
|---|---|---|
| | dialogs | words |
| recorded | 1984 | 505 K |
| transcribed | 1826 | 460 K |
| **German Scheduling** | | |
| | dialogs | words |
| recorded | 734 | 158 K |
| transcribed | 534 | 115 K |
| **Spanish Scheduling** | | |
| | dialogs | words |
| recorded | 340 | 79 K |
| transcribed | 256 | 70 K |
| ATIS3 | | |
| transcribed | n./a. | 250 K |

Table 1: Comparison of Databases (as of December 1994)

Table 1 summarizes the current status of data collection. Since Scheduling utterances typically consist of more than one sentence, there is already more data available for English Scheduling than ATIS [1]. More data collection will establish databases in size at least comparable to ATIS for all languages.

In Spanish, we have explored two different data collection scenarios: To allow only one person to speak at a time the *push-to-talk* scenario requires the speaker to push a button while talking to the system. The *cross-talk* scenario allows speakers to speak simultaneously without push button. The speech of each dialog partner is recorded on separate channels.

### 2.2. System Overview

The main system modules are speech recognition, parsing, discourse processing, and generation. Each module is lan-

[1] The about 18000 utterances in English Scheduling correspond to some 30000 sentences.

guage-independent in the sense that it consists of a general processor that applies independently specified knowledge about different languages.

The recognition module decodes the speech in the source language into a list of sentence candidates, represented either as a word lattice or Nbest list. At the core of the machine translation components is a language independent representation of the meaning, which is extracted from the recognizer output by the parsing module. As last step, the final language independent representation is sent to the generator to be translated in any of the target languages. Figure 1 shows the system architecture.

After parsing, a discourse processor can be used to put the current utterance in the context of previous utterances, opening possibilities to integrate the speech and natural language processing compenents of the system to resolve parsing ambiguities and dynamically adapt the vocabulary and language model of the recognizer based on the current discourse state.

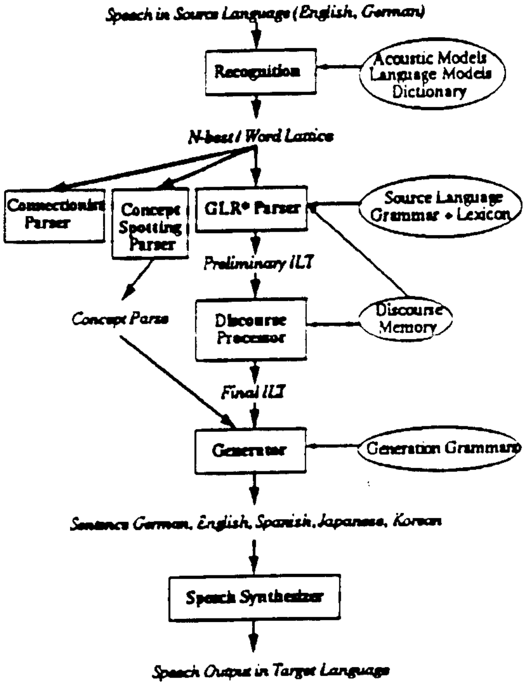


Figure 1: System Diagram

We explore several approaches for the main processes. For example, we are experimenting with TDNN, MS-TDNN [3], MLP, LVQ [4], and HMM's [5, 12] for acoustic modeling; n-grams, word clustering, and automatic phrase detection for language modeling [6]; statistically trained skipping parsing [7, 8], neural net parsing [9] and concept spotting parsing [10] for extracting the meaning; and statistical models as well as plan inferencing for identification of the discourse state [11]. This multi-strategy approach should lead to improved performance with appropriate weighting of the output from each strategy.

## 2.3. Recognition Performance Analysis

The baseline JANUS-2 recognizer can be described as follows:

- *Preprocessing*: LDA on melscale fourier spectrum and additional acoustic features (power, silence)
- *Acoustic modeling*: LVQ-2 or phonetically tied SCHMM, no cross–word triphones, explicit noise models
- *Decoder*: Viterbi search as first pass, followed by a word-dependent Nbest search, standard word bigram language model, word lattice output

Current recognition results on the English, German and Spanish Spontaneous Scheduling Task (ESST, GSST, SSST) can be seen in table 2.

| | ESST | GSST | SSST |
|---|---|---|---|
| Word Accuracy | 66% | 72% | 61% |

Table 2: JANUS-2 baseline recognition performance

The low absolute recognition accuracies are due to the challenging nature of human–human spontaneous speech. In the official evaluation of the German VERBMOBIL project on the GSST task, the JANUS-2 decoder outperformed all other participating systems. In addition, recent evaluations on the Switchboard task confirm that human–human dialogs are much more difficult to recognize than human–machine spontaneous speech (like ATIS). Participating systems achieved word accuracies between 30% and 50%.

Analysis shows that human-human dialogs (like Scheduling or Switchboard) are more difficult to recognize than human-machine dialogs (e.g. ATIS). Perplexities lie between 35 and 90 for ESST, SSST and GSST, and somwhat over 100 for Switchboard. Additionally, human-human dialogs are significantly more disfluent [8]. Large variations in speaking rates and strong coarticulation between words contribute significantly to the difficulty of recognizing human-human spontaneous speech.

## 3. Improving the Recognition Component

We describe efforts to improve the recognition component along its major knowledge sources acoustic models [12], dictionary [13] and language models [14].

## 3.1. Data–Driven Codebook Adaptation

We developed methods aimed at automatic optimization of the number of parameters for the semi-continuous phonetically tied HMM used in JANUS-2. Usually, a fixed number of codebook vectors is assigned to each of the phonemes. However, as the available training data differs between phonemes and the size of the feature space phonemes cover varies greatly, constant codebook size leads to suboptimal allocation of resources.

We have therefore suggested [12] to adapt the codebook size of each phoneme according to the amount and the distribution of the training data, similar to [15]. During training, the size of the codebook is incrementally increased. Some quality criterion determines when to stop the process of increasing the codebook. We compared a *variance* criterion based on the average distance between data points and their nearest codebook vector with a *prediction* criterion which tries to capture how well the modeling of the recognizer can predict unseen data.

| Model | Codebook Size | Word Accuracy |
|---|---|---|
| baseline | 4600 | 66.9% |
| variance | 4201 | 69.9% |
| prediction | 1677 | 67.8% |

Table 3: Results for Codebook Adaptation (GSST)

Table 3 compares recognition accuracies and codebook sizes of the baseline models, with models automatically adapted using the variance and prediction criterion. As can be seen, codebook adaptation leads to significant error reduction if the same number of parameters is used.The number of parameters can be reduced by 40% with still better performance than the baseline system.

## 3.2. Dictionary Learning

Due to the enormous variability in spontaneous human–human dialogs creating adequate dictionaries with alternative pronunciations is crucial [16]. However, hand tuning and modifying dictionaries is time consuming and labor intensive. Pronunciations of a word should be chosen according to their frequency. Modifications of the dictionary should not lead to higher phonetic confusability after retraining. Therefore we have proposed [13] a data-driven approach to improve existing dictionaries and automatically add new words and pronunciation variants whenever needed.

The learning algorithm requires transcripts for the whole training set and a phoneme confusability matrix of the speech recognizer used. First, phonetic transcriptions for all appearances of each word are generated by help of a phoneme recognizer. Then, variants which are infrequent or which would lead to erroneous training of confusable phonemes are eleminated. Finally, the acoustic models are retrained allowing for the newly aquired pronunciations variants.

As can be seen in table 4, our algorithm for adapting and adding phonetic transcriptions to a dictionary improves the recognition accuracy of the decoder significantly and leads to performance that is comparable to the context dependent results (cf. table 2). The baseline decoder for these experiments uses 69 context independent phoneme models. Evaluation using context dependent models is in progress.

| Dictionary | Word Accuracy |
|---|---|
| baseline | 61.7% |
| adapted | 65.6% |

Table 4: Results Dictionary Learning (GSST)

## 3.3. Morpheme Based Language Models

Based on our scheduling databases we noticed that in morphologically rich languages such as German and Spanish, dictionaries grow much faster with increasing database size, compared to English (cf. figure 2). This is due to the large number of inflections and compound words. One way to limit this growth with increasing dictionary sizes is to use other base units than words.

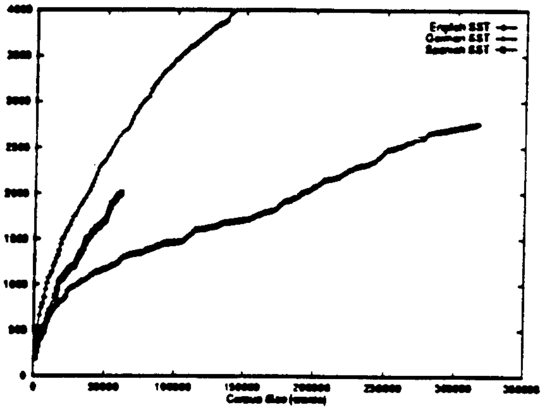


Figure 2: Vocabulary Growth

We compared three different decomposition methods:

- strictly *morpheme* based decomposition, e.g. weggehen (to go away) ⟶ weg–geh–en, Spracherkennung (speech recognition) ⟶ Sprach–er–kenn–ung
- decomposition in *rootforms*, e.g. weggehen (to go away) ⟶ weggeh@, Spracherkennung (speech recognition) ⟶ Spracherkenn@

- combination of strictly morpheme based decomposition and root forms

Table 5 shows dictionary size, bigram perplexity and recognition accuracy using the respective decomposition method, based on 250 GSST dialogs. As can be seen, all decomposition methods significantly reduce vocabulary size and perplexity. The impact on recognition accuracy is still small. This may be due to the fact that the acoustic modeling suffers from smaller units and thus deteriorate the gain in the language model. In a real interface, however, this reduction in vocabulary growth leads to a reduction of new words. Further research will focus on finding more efficient and acoustically less confusable decompositions automatically, and also test the impact on translation.

| | Dictionary | Perplexity | Accuracy |
|---|---|---|---|
| Baseline | 3821 | 88 | 64.7% |
| Morphemes | 2391 | 46 | 65.4% |
| Root Forms | 3205 | 79 | 63.5% |
| Combined | 2998 | 59 | 65.1% |

Table 5: Comparison of Decomposition Methods (GSST)

## 4. Concept Based Speech Translation

We have developed a robust translation system based on the information structures inherent in the appointment scheduling task being performed, described in detail elsewhere [10]. The basic premise is that the structure of the information conveyed is largely independent of the language used to encode it. Our system tries to model the information structures in a task and the way these structures are realized in words in various languages. This system is an extension of the Phoenix Spoken Language System [18]. It uses the Phoenix parser to parse input into slots in semantic frames, and then uses these frames to generate output in the target language.

### 4.1. The Parser

Unlike individual words, semantic units used in a task domain are not language specific. Based on transcripts of scheduling dialogs, we have developed a set of fundamental semantic units in our parse which represent the different concepts a speaker would use. For instance, a typical *temporal* token could have *date* as subtoken, which could in turn consist of *month* and *day* subtokens. The *temporal* could be part of a statement of unavailability.

In contrast to previous speech translation systems, we presently don't perform syntactic analysis. Speaker utterances, as decoded by the recognizer, are parsed into semantic chunks which are concatenated without grammatical rules.

```
Original utterance:
  THAT SATURDAY I'M NOT SURE ABOUT BUT YOU SAID
    YOU MAY BE BACK IF YOU THINK YOU'LL BE BACK
    THE THIS SUNDAY THE TWENTY EIGHTH I COULD SEE
    YOU AFTER ELEVEN AM ON THAT IF YOU'RE BACK

Translated:
```

*Saturday that's not so good for me Sunday the twenty eighth works for me after eleven a.m. (ENGLISH)*

*El sábado no me va demasiado bien pero el domingo veintiocho me va bien después de las once de la mañana. (SPANISH)*

*Samstag könnte ich nur zur Not aber Sonntag der Achtundzwanzigste geht bei mir ganz gut nach elf Uhr morgens. (GERMAN)*

Figure 3: Translation Example

This approach is particularly well suited to parsing spontaneous speech, which is often ungrammatical and subject to recognition errors. This approach is more robust than requiring well-formed input and reliance on syntactic cues provided by short function words such as articles and prepositions.

### 4.2. The Generator

The generation component of the system is a simple left-to-right processing of the parsed text. The translation grammar consists of a set of target-language phrasings for each token, including lookup tables for variables like numbers and days of the week. When a lowest-level token is reached in tracing through the parse, a target-language representation is created by replacing tokens with templates for the parent token, according to the translation grammar. The result is a meaningful, although terse translation, which emphasizes communicating the main point of an utterance. An examples is illustrated in figure 3.

### 4.3. Results

We have implemented this system for bi-directional translation between English, German and Spanish in our scheduling task. Table 4 shows the performance of parser and subsequent generator on transcribed data. Evaluation of the system based on speech decoded by the JANUS-2 recognizer is still underway.

| | Parsed from | | Translated into |
|---|---|---|---|
| | token | utterance | utterance |
| English | 95.6% | 90.0% | 90.2% |
| German | 92.4 | 89.6 | 87.3 |
| Spanish | 88.8 | 58.3 | 82.2 |

Figure 4: End–to–End evaluation on transcribed data

One disadvantage of this approach is the telegraphic and repetitive nature of the translations. This could be overcome by providing multiple translation options for individual tokens in the target-language module, different levels of politeness, etc. However at present we feel that it is sufficient for intelligible communcation.

## 5. GLR* Parser

In addition to the concept based Phoenix parser we pursue GLR* as robust extension of the Generalized LR Parser. It attempts to find maximal subsets of the input that are parsable, skipping over unrecognizable parts of the input sentence [7]. By means of a semantic grammar GLR* parses input sentences into an interlingua text (ILT) as language independent representation of the meaning of the input sentence, described in more detail elsewhere (e.g. [8]).

Compared to Phoenix parses the ILT generated by GLR* offers greater level of detail and more specificity, e.g. different speaker attitudes and levels of politeness. Thus, translation based on ILTs is more natural, overcoming the telegraphic and terse nature of concept based translation.

A drawback of GLR* was that it expected input segmented into sentences for efficiency reasons. However, typical Scheduling utterances consist of 2-3 sentences. To integrate the parser with the speech decoder, we developed methods which extend the parsing capabilities from single sentences to multi-sentence utterances. We extended the grammar with a high-level rule that allows the input utterance to be analyzed as a concatenation of several sentences and developed two methods to constrain the number of sentence breaks that are considered by the parser. The first is a heuristic which prunes out all parses that are not minimal in the number of sentences. The second is a statistical method to disregard potential sentence breaking points that are statistically unlikely.

For the English analysis grammar, time efficiency thus improved by about 30%. As an additional benefit, the parse quality improved because strange sentence breaks are rejected in favor of a more reasonable location.

## 6. Handling Unreliability

Although research has boosted performance of speech recognition and spoken language translation technology, recognition and translation errors will persist. To build a system for use in real applications we need repair methods to recover from errors in a graceful and unobstrusive way. We have developed a speech interface for repairing *recognition* errors by simply respeaking or spelling a misrecognized section of an utterance. While much speech "repair" work has focused on repairs within a single spoken utterance [19], we are concerned with the interactive repair of errorful recognizer hypotheses [20].

### 6.1. Identifying Errors

To be able to repair an error its location has to be determined first. We pursue two strategies to identify misrecognitions as subpieces of the initial recognizer hypothesis.

The *automatic subpiece location* technique requires the user to respeak only the errorful subsection of the (primary) utterance. This (secondary) utterance is decoded using a vocabulary and language model limited to substrings of the initial erroneous hypothesis. Thus, the decoding identifies the respoken section in the hypothesis. Preliminary testing showed that the method works poorly if the subpiece to be located is only one or two words long. However, this drawback is not severe since humans tend to respeak a few words around the error.

A second technique uses *confidence measures* to determine for each word in the recognizer hypothesis whether it was misrecognized. First, we applied a technique similar to Ward [21], which turns the score for each word obtained during decoding into a confidence measure by normalizing the score and using a Bayesian updating technique based on histograms of the normalized score for correct and misrecognized words. Since we found this not to work well on our English scheduling task, we are currently developing different methods to compute confidence measures based on decoder, language model and parser scores.

### 6.2. Robust Speech Repair

After locating and highlighting erroneous sections in the recognizer hypothesis misrecognitions are corrected.

The *spoken hypothesis correction* method uses Nbest lists for both the initial utterance and the respoken section. The Nbest for the highlighted section of the initial utterance is rescored using scores from decoding the secondary utterance. Depending on the quality of the Nbest lists, most misrecognitions can be corrected.

The *spelling hypothesis correction* method requires the user to spell the highlighted erroneous section. A spelling recognizer decodes the spelled sequence of letters. By means of a language model we restrict the sequence of letters to alternatives found among the Nbest from the located section.

To date, we have evaluated our methods over sentences from the Resource Management task. Table 6 shows the improvements in sentence accuracy, based on recordings from one speaker of the February and October 1989 test data. We selected a subset of erroneous utterances; therefore the accuracy of the baseline system is significantly lower than the 94% performance our system achieves on the whole test set. The results indicate that repeating or spelling a misrecognized subsection of an utterance can be an effective way to repair recognition utterances.

| | |
|---|---|
| No Repair (baseline) | 63.1% |
| Respeak | 83.8% |
| Spell | 88.5% |
| Respeak + Spell | 89.9% |

Table 6: Improvement of Sentence Accuracy by Repair

## 7. Conclusions

We have made significant advances towards building a multi lingual translation system for spontaneous human–human dialogs. Beyond speech recognition of spontaneous speech JANUS provides a framework to investigate important areas like robust parsing, machine translation of spoken language and developing methods to recover from recognition and parsing errors. To achieve acceptance in real applications, we have to embed the spoken language technology in a sensible and useful user interface that is carefully designed around human factors and common needs. To be flexible and robust, such interfaces should not only recognize speech but also recognize other communication modalities, provide freedom from headset and push-buttons, allow for graceful recovery from errors and miscommunications, know what they don't know, and model what the user does or doesn't know [23].

## 8. Acknowledgements

This research was funded in part by grants from the Advanced Research Project Agency and the German Ministry of Science and Technology (BMFT) under project Verbmobil. We also gratefully acknoledge support by ATR Interpreting Telecommunications Research Labs of Japan. The views and conclusions contained in this document are those of the authors.

## References

1. L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, M. Woszczyna: *Testing Generality in JANUS: A Multi-Lingual Speech to Speech Translation System.* Proc. ICASSP 92, vol. 1, pp. 209-212
2. M. Woszczyna, , N. Coccaro, A. Eisele, A. Lavie, A.-E. McNair, T. Polzin, I. Rogina, C. Pennstein-Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki–Waibel, A. Waibel, W. Ward: *Recent Advances in JANUS: A Speech Translation System,* DARPA Speech and Natural Language Workshop 1993, session 6 - MT
3. H. Hild and A. Waibel: *Connected Letter Recognition with a Multi–State Time Delay Neural Network,* Neural Information Processing Systems (NIPS-5), Morgan Kaufman
4. O. Schmidbauer and J. Tebelskis: *An LVQ based Reference Model for Speaker–Adaptive Speech Recognition,* Proc. ICASSP 92, Vol. 1, pp. 441-445
5. I. Rogina and A. Waibel: *Learning State–Dependant Stream Weights for Multi–Codebook HMM Speech Recognition Systems,* Proc. ICASSP 94
6. B. Suhm and A. Waibel: *Towards Better Language Models for Spontaneous Speech,* ICSLP 94, Vol. 2, pp. 831-834
7. A. Lavie and M. Tomita: *GLR* - An Efficient Noise-skipping Parsing Algorithm for Context-free Grammars,* Proceedings of Third International Workshop on Parsing Technologies, 1993, pp. 123–134
8. B. Suhm, L. Levin, N. Coccaro, J. Carbonell, K. Horiguchi, R. Isotani, A. Lavie, L. Mayfield, C. Pennstein-Rosé, C. Van Ess-Dykema and A. Waibel: *Speech–Language Integration in a Multi–Lingual Speech Translation System,* Workshop on Integration of Natual Language and Speech Processing, AAAI-94, Seattle
9. F.-D. Buø, T.-S. Polzin and A. Waibel: *Learning Complex Output Representations in Connectionist Parsing of Spontaneous Speech,* Proc. ICASSP 94, Vol. 1, pp. 365-368
10. L. Mayfield, M. Gavalda, W. Ward and A. Waibel: *Concept Based Speech Translation,* to appear in Proc. ICASSP 95
11. Carolyn Penstein Rosé, Alex Waibel: *Recovering From Parser Failures: A Hybrid Statistical/Symbolic Approach,* to appear in "The Balancing Act: Combining Symbolic and Statistical Approaches to Language" workshop at the 32nd Annual Meeting of the ACL, 1994
12. T. Kemp: *Data-Driven Codebook Adaptation in phonetically tied SCHMMS,* to appear in Proc. ICASSP 95
13. T. Sloboda: *Dictionary Learning: Performance through Consistency,* to appear in Proc. ICASSP 95
14. P. Geutner: *Using Morphology towards better Large Vocabulary Speech Recognition Systems,* to appear in Proc. ICASSP 95
15. U. Bodenhausen: *Automatic Structuring of Neural Networks for Spatio-Temporal Real-World Applications,* Ph.D thesis, University of Karlsruhe, June 1994
16. J.-L. Gauvin, L.-F. Lamel, G. Adda and M. Adda-Decker: *The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task,* Proc. ICASSP 94, vol. 1, pp. 557-560
17. R. Kneser and H. Ney: *Improved Clustering Techniques for Class-Based Statistical Language Models,* EUROSPEECH 93, Berlin, Vol. 2, pp. 973-976
18. W. Ward: *Understanding Spontaneous Speech: The Phoenix System,* IEEE International Conference on Acoustics, Speech and Signal Processing, 1991, Vol. 1, pp. 365–367
19. C. Nakatani and J. Hirschberg: *A Speech–First Model for Repair Identification in Spoken Language Systems,* in Proceedings of the ARPA Workshop on Human Language Technology, March 1993
20. A.-E. McNair and A. Waibel: *Improving Recognizer Acceptance through Robust, Natural Speech Repair* ICSLP 94, Vol. 3., pp. 1299-1303
21. S.-R. Young and W. Ward: *Learning New Words from Spontaneous Speech,* Proc. ICASSP 93, Vol. 2, pp. 590–591
22. N. Yankelovich, G.-A. Levow and M. Marx: *Designing SpeechActs: Issues in Speech User Interfaces,* to be presented at CHI 95, Denver
23. M.-T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier and A. Waibel: *Multimodal Learning Interfaces,* to be presented at Spoken Language Technology Workshop 1995, Austin

# Machine-Aided Voice Translation (MAVT) Advanced Development Model*

Christine A. Montgomery, Ph.D.
Bonnie Glover Stalls, Ph.D.
Robert E. Stumberger
Naicong Li, Ph.D.
Robert S. Belvin
Alfredo R. Arnaiz
Susan H. Litenatsky

Machine-Aided Voice Translation (MAVT) is a development begun in 1990 for a spoken language translation prototype whose primary use is to assist Air Force personnel in interacting with speakers of foreign languages. The Phase I development resulted in a speaker-independent, continuous speech, medium vocabulary translation prototype for English => Spanish => English, which was installed at Rome Laboratory in 1992. The presentation and demonstration show the second phase of MAVT development -- which extends the foreign language repertoire of the system to include Arabic and Russian, and is based on an interlingual design. The MAVT system runs on any Sun workstation with 16-bit audio, and is written in C++ and Prolog. Speech recognition for all four languages is via Entropic Research Laboratory's HMM Tool Kit (HTK) software, and speech synthesis for English and Spanish is provided by Entropic's TrueTalk, licensed from AT&T. In the current version of the MAVT system, speech generation for Arabic and Russian is via digital audio playback.

Contact: Dr. Christine A. Montgomery
Language Systems, Inc.(LSI)
6269 Variel Avenue, Suite F
Woodland Hills, CA 91367
(818) 703-5034; FAX: (818) 703-5902
e-mail: chris@lsi.com

* This work is supported by Air Force Materiel Command, Rome Laboratory/IRAA, Griffiss Air Force Base, NY, under Contract No. F30602-93-C-0098.

Reprinted from the 5th Annual IEEE
Dual-Use Technologies and Applications
Conference (May 22 - 25, 1995: 112-118)

# MACHINE-AIDED VOICE TRANSLATION (MAVT): ADVANCED DEVELOPMENT MODEL*

Christine A. Montgomery, Bonnie Glover Stalls, Robert S. Belvin,
Alfredo R. Arnaiz, Robert E. Stumberger, Naicong Li, Susan Hirsh Litenatsky
chris,glover,res,naicong,robin,arnaiz,hirsh}@lsi.com
Language Systems, Inc.

## Abstract

Machine-Aided Voice Translation (MAVT) is a development begun in 1990 for a spoken language translation prototype whose primary use is to assist Air Force interrogation personnel in interacting with speakers of foreign languages. A significant potential use of the MAVT prototype is to provide similar support for law enforcement personnel, who have shown considerable interest in the development. The paper describes the second phase of MAVT development – which will result in a speaker-independent, continuous speech, multilingual translation prototype for English ⇒ Spanish|Arabic|Russian ⇒ English.

## 1 Introduction

Machine-Aided Voice Translation (MAVT) is a development begun in 1990 under contract to Rome Laboratory, AFMC, for a spoken language translation prototype to assist Air Force personnel in interacting with speakers of foreign languages. The initial phase of the project, which concluded in 1992, resulted in the development of a speaker-independent continuous speech, translation system for English ⇒ Spanish ⇒ English, using a vocabulary of about 500 words. An overview of the system as well as a summary of evaluation results are given in [1].

This paper describes the Phase II MAVT ADM system (Figure 1), which provides voice input and output for English ⇒ Spanish|Arabic|Russian ⇒ English, with a planned vocabulary of approximately 1,000 words per language. Like the Phase I system, the current system is comprised of three subsystems: a speech recognition system, a natural language processing system, and speech generators. Speaker-independent, continuous speech recognition is accomplished via Entropic's HMM Toolkit, while speech synthesis for English and Spanish utilizes Entropic's *TrueTalk*$^{tm}$, licensed from AT&T. (Generators for Arabic and Russian are still under negotiation at this time.) As in the Phase I system, natural language understanding and translation generation is achieved via LSI's DBG natural language processing system, which has been extended to incorporate a language-independent translation component that integrates predicate representations based on Jackendoff's Lexical Conceptual Structures (henceforth LCS) [2],[3] with DBG's frames and lexicon [4]. These three subsystems are briefly described in the following sections.

## 2 The DBG Natural Language Processing System

LSI's DBG system has served as the NLP engine for a variety of text understanding applications, focusing on information extraction for data base generation (from which the acronym DBG is derived) for a range of different types of text, and message fusion, based on a large sample of transcribed radiotelephone traffic. The components of the DBG system as configured for these applications include modules for lexical lookup and morphological analysis, full syntactic and semantic analysis, and discourse or text-level analysis. The analyzed content of a text is represented as a set of interconnected frame structures called templates, which reflect the entities and events described

*The work reported in this paper is supported by AFMC, Rome Laboratory/IRAA, Griffiss Air Force Base, NY, under Contract No. F30602-93-C-0098. Earlier work was supported under Contract No. F30602-90-C-0058.

## MACHINE - AIDED VOICE TRANSLATION

## ADVANCED DEVELOPMENT MODEL

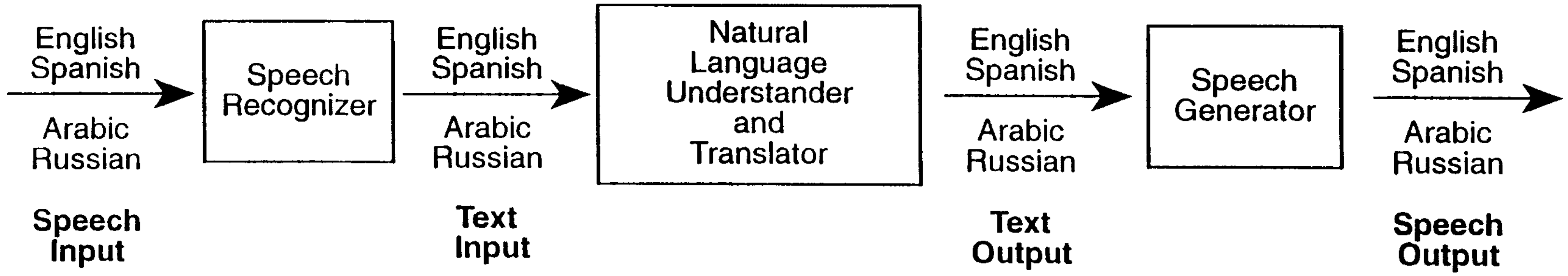


Figure 1. Version 1.5 MAVT ADM System Diagram

in a source text.

For the MAVT application, modules were added to generate the target language text. In the Phase I MAVT development, a direct transfer strategy was used to achieve translation, although many of the components were designed for multilingual use. In the current MAVT development, we have adopted an interlingual approach to translation. Much of the extension of the DBG system for the MAVT project has necessarily focused on the multilingual capabilities of the system. In the first phase of the project, the DBG system already had in place a multilingual syntactic parser that was used for Spanish and English. An updated version of this parser will be used to parse Arabic and Russian as well. DBG also produces, as output of the understanding phase of processing, a knowledge representation of the sentence. This knowledge representation is an application-independent data structure of related event and entity frames based on the predicates and arguments of the sentence, as well as on an underlying frame-based concept hierarchy. These frames, called *templates* in the DBG system, represent the knowledge contained in a sentence. On the basis of this structure, which is the end product of analysis of the source language (hereafter SL) sentence, the target language (TL) lexical items are selected, and generation processing is applied to construct a translation of the sentence.

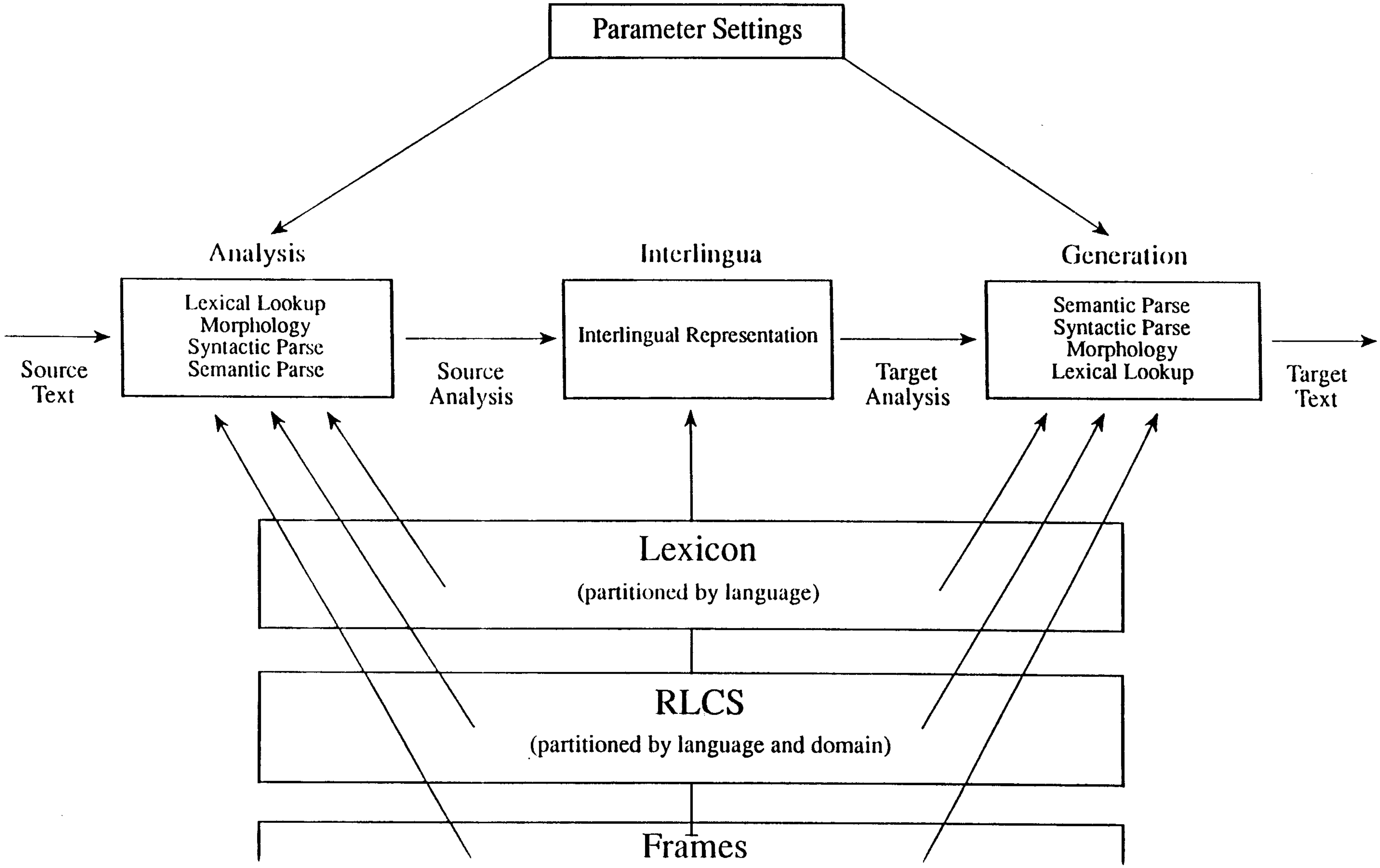
The DBG knowledge representation thus functions as an intermediate or *interlingual* (henceforth, IL) construct. An *IL* approach does not not rely on direct transfer or direct links between languages but requires a language-independent representation of the data, which can then be used to translate the sentence into any language that the system can handle. The IL approach thus eliminates the need to develop a separate, direct interface between every potential source-target language pair because each language need only interface with the language-independent IL representation.

From the commencement of the MAVT project, including the phase I development LSI's approach has been *interlingual* in that it assumes that the selection of lexical items in the TL should be based on links to an intermediate structure, rather than on direct or hard links between words in the source and target languages. In phase I, this was realized insofar as words corresponding to the same basic meaning in each language were linked to common concept nodes in the frame-based knowledge hierarchy. These links are present in each event and entity template in the knowledge representation.

For some lexical categories, e.g., nouns, this works well. But where cross-category relations and compositional semantics are important, as in verb phrases, which express predicate-argument relations, the lexical properties are much more complex. In a multilingual system, incorporating lexical-semantic information for the words associated with a given concept for all of the different languages into the concept hierarchy would greatly increase the complexity of the hierarchy. A limitation of using links to the concept hierarchy as the only intermediary, then, is that the concept hierarchy primarily represents meaning relations between concepts of the same category rather than representing the unique properties of the meanings of the individual words associated with those concepts, or the meaning relations and structural requirements of the words in sentences. A great deal of syntactic and semantic checking still remains to be done to determine whether a potential TL word is compatible with the meaning and structural requirements of the TL sentence. Thus, in our phase II development (the ADM phase), we determined it was highly desirable to construct an IL representation which could rely on some other knowledge source, beyond just the frame-based knowlege hierarchy. The emergent theory of Lexical-Conceptual Structures was determined to be highly appropriate as a means of encoding the additional knowledge representation required. These structures, when combined with DBG's existing interlingual characteristics, have proven to be exactly the link needed to create what we deemed was an appropriately robust IL representation.

The DBG system has a modular design, wherein text is analyzed in progressive stages. The output of each stage of processing is a data structure that then serves as input to the following processing stage. As illustrated in figure 2, there are four stages of SL analysis of a sentence that precede the IL template representation; the IL representation is then followed by four stages of TL generation. The four stages of SL analysis are: a)lexical identification, b) morphological analysis, c) syntactic parsing, and d) semantic parsing. The four stages of TL generation mirror in part the SL analysis: they are w) lexical selection, x) semantic parsing, y) syntactic parsing, and z) morphological inflection (see Figure 2; the acronym RLCS stands for "Root Lexical-Conceptual Structure", that

# MAVT-ADM : Language Translation



## Instantiated Interlingual Templates

### English Source Semantic Parse
[The tanks fired at the enemy]

| | | |
|---|---|---|
| mainpred (1.0) | = | INDEX (1.1) |
| utt. type (1.1) | = | declarative |
| predicate (1.1) | = | fire |
| slcs (1.1) | = | E[CAUSE(T[1.2],<br>E[GO_loc(T[<projectile>]<br>P[TO(T[1.3])])])] |
| tense (1.1) | = | past |
| voice (1.1) | = | active |
| aspect (1.1) | = | perfective |
| mood (1.1) | = | indicative |
| modality (1.1) | = | neutral |
| polarity (1.1) | = | positive |
| ext-arg (1.1) | = | INDEX (1.2) |
| obl arg (1.1) | = | INDEX (1.3) |
| noun (1.2) | = | tank |
| person (1.2) | = | 3rd |
| gender (1.2) | = | zero |
| number (1.2) | = | plural |
| class (1.2) | = | inanimate |
| determiner (1.2) | = | the |
| preposition (1.3) | = | at |
| prep object (1.3) | = | INDEX (1.4) |
| noun (1.4) | = | enemy |
| person (1.4) | = | 3rd |
| gender (1.4) | = | zero |
| number (1.4) | = | singular |
| class (1.4) | = | human collective |
| determiner (1.4) | = | the |

| | | |
|---|---|---|
| EVENT | | report [1] |
| class | : | meta |
| application | : | mavt-adm |
| domain | : | domestic mission |
| corpus | : | example |
| date | : | 13 apr 1994 |
| social situation | : | formal |
| genre | : | interrogation screening |
| event | : | [1.1] |

| | | |
|---|---|---|
| EVENT | | [1.1] |
| verb | : | fire |
| clcs | : | E[CAUSE(T[1.1.1],<br>E[GO_loc(T[<projectile>],<br>P[TO(T[1.1.2])])])] |
| text status | : | main clause |
| utterance type | : | assertion |
| discourse status | : | foreground |
| class | : | primary |
| status | : | critical |
| time | : | precedes [1] |
| voice | : | active |
| aspect | : | perfective |
| modality | : | neutral |
| polarity | : | positive |
| arg1 | : | [1.1.1] |
| arg2 | : | [1.1.2] |

| | | |
|---|---|---|
| ENTITY | | [1.1.1] |
| nucleus | : | tank |
| class | : | military vehicle |
| number | : | plural |
| sex | : | zero |
| person | : | 3rd |
| definiteness | : | definite |

| | | |
|---|---|---|
| ENTITY | | [1.1.2] |
| nucleus | : | enemy |
| class | : | human collective |
| number | : | singular |
| sex | : | zero |
| person | : | 3rd |
| definiteness | : | definite |

### Spanish Target Semantic Parse
[Los tanques le dispararon al enemigo.]

| | | |
|---|---|---|
| mainpred (1.0) | = | INDEX (1.1) |
| utt. type (1.1) | = | declarative |
| predicate (1.1) | = | disparar |
| slcs (1.1) | = | E[CAUSE(T[1.2],<br>E[GO_loc(T[<projectile>],<br>P[TO(T[1.3])])])] |
| tense (1.1) | = | past |
| voice (1.1) | = | active |
| aspect (1.1) | = | perfective |
| mood (1.1) | = | indicative |
| modality (1.1) | = | neutral |
| polarity (1.1) | = | positive |
| ext-arg (1.1) | = | INDEX (1.2) |
| dat-arg (1.1) | = | INDEX (1.3) |
| noun (1.2) | = | tanque |
| person (1.2) | = | 3rd |
| gender (1.2) | = | masculine |
| number (1.2) | = | plural |
| class (1.2) | = | military vehicle |
| determiner (1.2) | = | el |
| preposition (1.3) | = | a |
| prep object (1.3) | = | INDEX (1.4) |
| noun (1.4) | = | enemigo |
| person (1.4) | = | 3rd |
| gender (1.4) | = | masculine |
| number (1.4) | = | singular |
| class (1.4) | = | human collective |
| determiner (1.4) | = | el |

**Figure 3** (E = Event, T = Thing, P = Path).

is, the form of the LCS which is stored with the lexical root in the lexicon).

Stages a,b) and z) are mirror images of one another in that in a,b) inflected lexical items are analyzed to determine their lexical stems and morphological features, and in z) lexical stems are inflected based on the accompanying morphological features. Likewise, c) and y) are very similar in that in both the internal syntactic structure associated with the sentence is organized in a principle-based manner, using a binary-branching version of x-bar theory. The difference between c) and y) is that in c) the structure of the SL sentence is discovered based on lexical and morphological information derived from an actual sentence, whereas in y) the syntactic structure is being built based on a semantic outline of the proposed TL sentence.

At the heart of processing in the DBG translation system are the three intermediate stages: the SL semantic parse (d, above), the IL templates, and the TL semantic parse (x, above). These are where translation occurs and it is into these data structures that we have incorporated Jackendoff's LCS (as mentioned earlier). An LCS is a labeled bracketing, similar to a syntactic parse structure, but one wherein the constituents labels, predicates and arguments are semantically-based primitives, rather than syntactic and language-specific lexical items. The data structures at these three stages are essentially of the same type: sets of attribute-value pairs related to other pairs by means of indexing. This kind of structure allows the system to pass on actual sentence chunks, along with associated features of whatever type, e.g., morphological, semantic, pragmatic, in a homogeneous format. An actual example of the three intermediate stages is provided in figure 3. A detailed discussion of this innovative development is presented in our paper for the AMTA 94 conference [4].

## 3 ASR via HTK: an HMM Software Toolkit

The speech recognition component of MAVT-ADM is an HMM toolkit. Entropic Research Laboratory licenses this technology from the Cambridge University Technology Transfer Company, and is responsible for ongoing support of HTK and future enhancements. HTK allows flexible development and modification of speaker models (e.g., recognizers for different languages and applications) based on Hidden Markov Model (HMM) principles, for isolated, connected, or continuous speech recognition. The recognizer is syntax-driven, via a finite state grammar which is customized for a particular recognition task. In recent ARPA testing of speech recognition systems developed by ARPA contractors and others, the HTK-based system performed comparably with those of ARPA contractors on dictation tasks involving a 5,000 word vocabulary and a 20,000 word vocabulary derived from Wall Street Journal texts. On the 5,000 word task, the recognizer developed with HTK performed at 95% accuracy, performing at 87% for the complex 20,000 word dictation task. HTK is written in ANSI C, and runs on Sun, H-P, DEC, or SGI workstations under Unix.

In the initial demonstration version of the MAVT ADM, speaker-independent, continuous speech recognizers for a limited mission-oriented vocabulary have been developed for English, Latin American Spanish, Arabic, and Russian.

## 4 *TrueTalk*$^{tm}$ Text-to-Speech (TTS) Software

*TrueTalk*$^{tm}$ is an advanced software-only TTS system that converts digitized text into speech, with a word intelligibility rate of approximately 97%. Entropic licenses this technology from AT&T, where it has been in development over the past 10 years. *TrueTalk*$^{tm}$ features a variety of user controls, including pitch, word duration, intonation, and speaking rate. For English, *TrueTalk*$^{tm}$ uses a primary dictionary of 166,000 words, and a secondary dictionary to assist in accurate pronunciation of proper names, such as location designations. The Spanish vocabulary is of a comparable size. *TrueTalk*$^{tm}$ runs on Sun, H-P, or SGI workstations under Unix.

## References

[1] C. Montgomery, B.G. Stalls, R.E. Stumberger, N. Li, S. Walter, R. Belvin, and A. Arnaiz, "Machine-aided voice translation", *in Information Management Collection Processing & Distribution, Dual-Use Technologies & Applications Conference*, pp. 96–101. IEEE, 1993.

[2] R. Jackendoff, *Semantic Structures*, MIT Press, 1990.

[3] B.J. Dorr. *Machine Translation: A View from the Lexicon*. MIT Press. 1993.

[4] B.G. Stalls. R.S. Belvin. A.R. Arnaiz. C.A. Montgomery, and R.E. Stumberger. "An adaptation of lexical conceptual structure to multilingual processing in an existing text understanding system". *in Technology partnerships for crossing the language barrier*, pp. 106–113. AMTA. 1994.

## Automatic English-to-Korean Text Translation of Naval Operational Reports

Young-Suk Lee, Dinesh Tummala,
Stephanie Seneff, Cliff Weinstein, and Jack Lynch

The automatic English-to_korean text translation project in our group is based on the natural language understanding system TINA (S. Seneff, 1992) and the generation system GENESIS (J. Glass, J. Polifroni, and S. Seneff, 1994), which were developed under ARPA sponsorship by the Spoken Language Systems Group at the MIT Laboratory of Computer Science. The overall goal of the project is to produce machine translation of both text and speech for enhanced multilingual and multinational operations. This project has its origins in the CCLINC translation system (Tummala et al 1993). CCLINC is an automatic speech-to-speech translation system for limited-domain multilingual applications including English, French and Korean.

The MUC-II data, our source language data, consists of 105 naval messages, which feature incidents involving different platforms such as aircraft, surface ships, submarines, and land targets. The data contain linguistically challenging features such as numerous instances of coordination, complex sentences, multiple modifiers, and compound nouns. At the same time, the data have typical characteristics of free texts including ellipsis and misspelling. We have translated 206 sentences (out of 643 sentences), and built up an English/Korean bilingual lexicon containing 432 vocabulary items, which is easily reusable by other systems (including PC-based ones).

The system demonstrated runs on a SPARC 10 workstation. The Korean translation outputs are displayed on a 'hangul' window running on UNIX, and the Korean inputs are typed in 'hangul' emacs, a version of emacs customized to support Korean alphabets.

[Contact authors for references.]

Contact: Dr. Clifford Weinstein
MIT Lincoln Laboratory
244 Wood St., Rm. S4-131
Lexington, MA 02173-9108
(617) 981-7491; FAX: (617) 981-0186
e-mail: cjw@sst.ll.mit.edu

# CCLINC: System Architecture and Concept Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications[1]

*Dinesh Tummala, Stephanie Seneff* [2], *Douglas Paul* [3], *Clifford Weinstein, Dennis Yang*

Lincoln Laboratory, MIT
Lexington, Ma. 02173-9108

## Abstract

This paper describes CCLINC, a system architecture and concept demonstration for automatic speech-to-speech translation for limited-domain multilingual applications. The primary target application is the coalition battle management environment. CCLINC utilizes a Common Coalition Language (CCL) as a military interlingua. CCLINC is a speaker-independent system which translates spoken utterances in English into French or Korean. The current system has a vocabulary of around 700 words. The system architecture for CCLINC consists of a modular, multilingual structure including speech recognition, language understanding, language generation, and speech synthesis in each language. A key new feature of the system is the tight coupling of the speech recognition and language understanding modules. We summarize the architectures of the component systems and the interfaces between them, and present our preliminary performance results.

## 1. Introduction

This paper describes a system architecture and concept demonstration for automatic speech-to-speech translation for limited-domain multilingual applications. (Other speech-to-speech translation systems are described in [9, 10, 13].) The primary target application is enhanced communication among military forces in a multilingual coalition environment, where the translation utilizes a Common Coalition Language as a military interlingua. This interlingua is designed to allow representation of the meanings of the limited-domain communications among forces in a common format for transmission.

The system architecture (see Figure 1) for CCLINC consists of a modular, multilingual structure including speech recognition, language understanding, language generation, and speech synthesis in each language. The meaning representation is in the form of a semantic frame, which is transmitted over the Common Coalition Language network. The system design provides for verification of the system's understanding of each utterance to the originator, in a paraphrase in the originator's language, before transmission on the coalition network. Successful system operation depends on the ability to define a sufficiently constrained, but useful, vocabulary and grammar, so that a high percentage of input sentences can be successfully understood. This understanding would also provide the opportunity to carry out update and query of command and control databases via CCL, along with the translation for human communication.

The rest of the paper is organized as follows. First, we describe CCLINC, paying particular attention to the speech recognition and natural language components as well as the interface between these components. Then we describe the training and present and evaluate the results of our preliminary experiments. This is followed by a discussion of lessons learned. Finally, we give our future plans.

## 2. System Description

### 2.1 Overview

The preliminary implementation of the CCLINC system uses a version of the Lincoln stack-decoder-based HMM system for continuous speech recognition[7, 8], in conjunction with language understanding (TINA)[1, 11, 15] and language generation (GENESIS)[2] systems which have been ported from the Spoken Language Systems Group at the MIT Laboratory for Computer Science. The vocabulary, grammar, and semantics are based on a coalition brigade task and are defined based on consultation with Army personnel and others familiar with brigade communications, a specification of command and control message formats, and a limited set of transcribed brigade exercise communications. For instance, the system has knowledge of basic Army radio-telephone vocabulary (e.g., roger, break, etc.), Army radio-telephone protocols (e.g., user identification), and basic military terms (e.g., weapons as well as terms such as TOC [tactical operation center] and FLOT [forward line of troops]). The current working vocabulary is 692 words[4] and the domain includes 253 semantic categories in the brigade communications domain.

CCLINC currently handles many sentences of moderate linguistic complexity. In particular, CCLINC understands both the active and passive voice and numerous verb forms (e.g., present tense, past participle, present participle, and imperative). The current system deals with three languages, English, Korean, and French. It accepts English speech/text input only, and translates via CCL to Korean (Hangul) or French text. We are using a commercial text-to-speech system on the English paraphrases which are produced based on the semantic understanding. We have recently obtained but not yet integrated a Korean text-to-speech synthesizer.

[1] This work was sponsored by the Advanced Research Projects Agency. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

[2] Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3] Now with Dragon Systems Inc., 320 Nevada St., Newton, MA. 02160.

[4] Although all versions of CCLINC recognize 692 words, some versions do not have any meaningful training data for 171 of these words. We will have more to say about this in section 3.1.

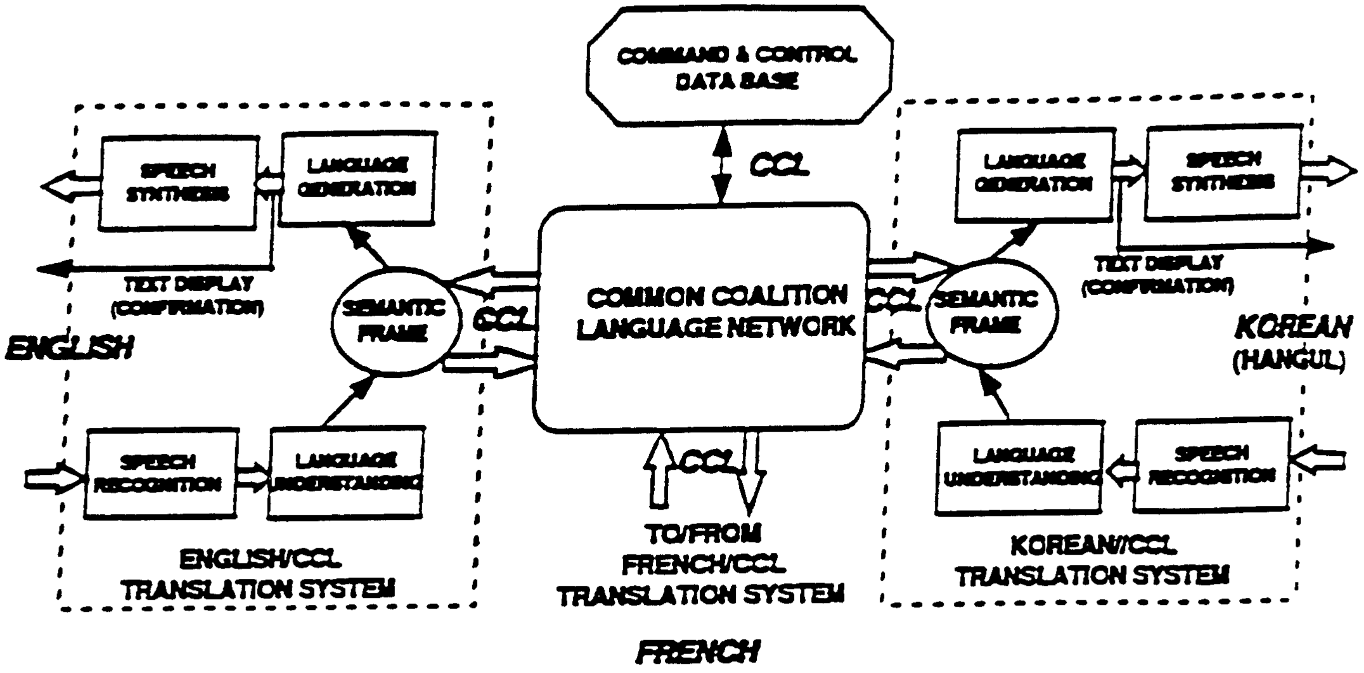


Figure 1: System structure for multilingual speech-to-speech translation.

We do have as yet a French speech synthesizer. Figure 2 shows an overview of CCLINC.

### 2.2 Speech Recognition

The preliminary CCLINC system uses Lincoln's large-vocabulary stack-decoder-based HMM in conjunction with a set of speaker-independent, trigram acoustic models[4, 8] and an augmented Carnegie Mellon Pronouncing Dictionary for speech recognition.

### 2.3 SR/NL Integration

The integration of continuous speech recognition (CSR) and natural language (NL) models has been an important part of this effort. We have implemented a new, tightly-coupled approach in which the TINA language model is integrated directly into the stack-based search[3]. For comparison, we have also implemented the type of decoupled approach in more general use in the ARPA community, where the 1-best or N-best CSR pipes its output into the language understanding module.[5] Thus, the recognizer runs in two different modes: a decoupled mode and a tightly-coupled mode, hereafter referred to as TINA-LM. In the decoupled mode, the recognizer is supported by a statistical language model; we have run experiments with a data-driven bigram backoff language model, a data-driven trigram backoff language model, and a TINA-generated bigram backoff language model. The TINA-generated bigram is created by expanding TINA's rules exhaustively to the terminals, multiplying out conditional probabilities along the way. In the tightly-coupled mode, TINA provides the sole linguistic support for the recognizer, proposing probabilities for each next word that is allowed by the grammar.

### 2.4 Machine (Text) Translation

The current CCLINC system uses TINA and GENESIS as its NL component (i.e., to perform machine or text translation). Machine translation systems vary along two major dimensions: basic approach (i.e., operation by statistical vs. symbolic/linguistic means) and depth of analysis (i.e., direct replacement, transfer, or interlingual)[7]. TINA/GENESIS is classified as a symbolic/linguistic, interlingual machine translation system within this framework.[6] TINA is based on a context-free grammar augmented with syntactic and semantic features[1, 11, 15]. The parser, with the aid of a morphological analyzer, produces a parse tree representation of the input sentence. This parse tree is then mapped to a semantic frame, which is the starting point for the language generation module, GENESIS.

GENESIS produces a paraphrase in the target language from the semantic frame[2]. The semantic frame is intended to capture the meaning of an utterance in a way that preserves the hierarchical dependencies in the utterance. Language generation is effected by the interaction of the language-independent, GENESIS engine with three language-specific modules. These modules are a lexicon, a set of message templates, and a set of rewrite rules. The main role of the lexicon is to specify the surface form of a semantic frame entry, including the construction of inflectional endings. The catalog of message templates determines the ordering of constituents in a sentence. The third module, the rewrite rules, captures phonotactic constraints and contractions. For instance, in French, "de les" is realized as "des."

Figure 3 and Figure 4 show the parse tree, semantic frame, and paraphrases produced by CCLINC for the sample sentence, "Request permission to defend hilltop echo." One

[5] At the current time, we only ran a 1-best CSR.

[6] Although TINA's rules are entered manually, TINA includes a probabilistic framework, along with an automatic training capability.

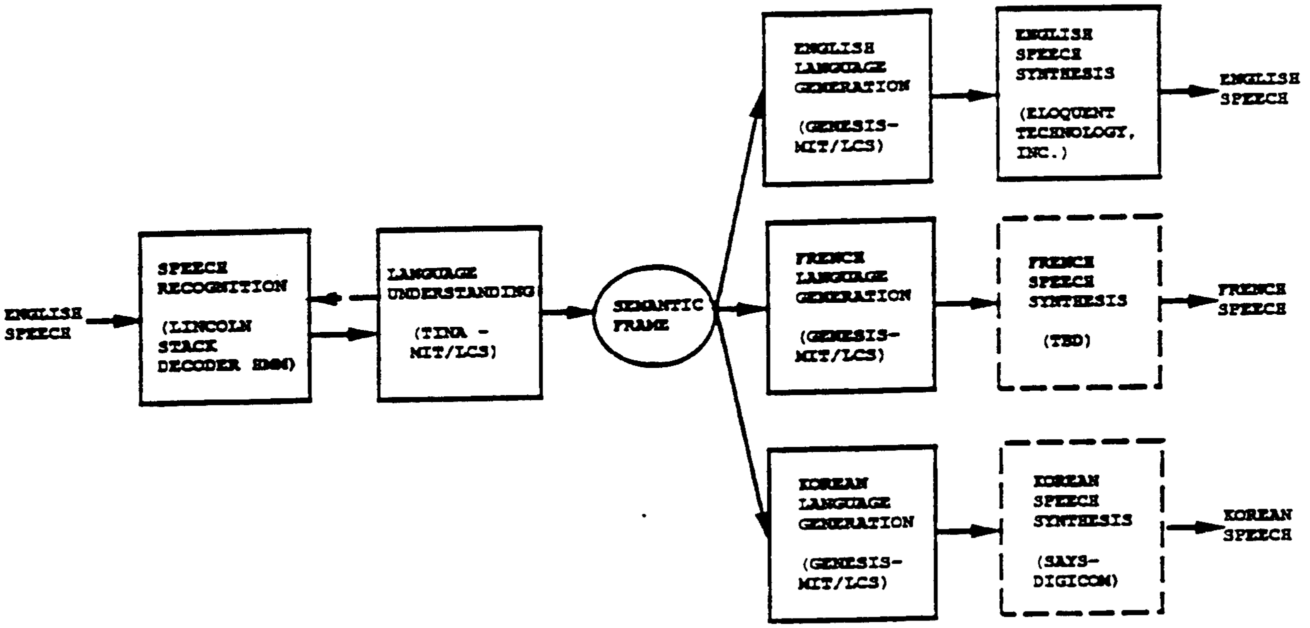


Figure 2: Process flow of CCLINC.

point to note in Figure 3 is the presence of syntactic categories near the root of the tree (i.e., statement, predicate, infinitive, etc.) and semantic categories near the leaves of the tree (i.e., fortify, the_location, etc.). Also note that the sentence which is translated in Figure 3 and in Figure 4 is a statement. A sentence in the coalition brigade domain is either a statement, command, callup (i.e., a sentence in which a user identifies himself), or reply (i.e., a subjectless phrase which may include, among other things, an opening remark such as "roger", a command and control message such as "sitrep," and/or a closing remark such as "over").

An English paraphrase of the sample sentence as well as translations in French and Korean appear in Figure 4. Note that the English paraphrase differs from the input sentence in two ways. First, we have inserted the subject "we." The input sentence does not contain an explicit subject. The implicit subject is "I" or "we." We arbitarily chose the plural "we" rather than the singular "I" as the subject. The second way in which the input sentence differs from its English paraphrase is in its choice of infinitive. The input sentence uses the word "defend" whereas the English paraphrase uses the word "fortify." The reason for this difference is that CCLINC generalizes the verb "defend." In fact, the verbs "defend," "fortify," and "strengthen" are all mapped to the same semantic category - the *fortify* category. The idea is to reduce the number of semantic objects known to the system (i.e., the number of lexical entries, the number of message templates, etc.) without losing meaning.

### 2.5 Text-to-Speech Synthesis

We have recently obtained, but not yet integrated, the Korean text-to-speech synthesizer "Says." "Says" is a product of Digicom. We do not have, as yet, a French speech synthesizer. On our English paraphrases, we are using a synthesizer developed by Eloquent Technology, Inc.

## 3. Training and Evaluation

### 3.1 Training

We are currently using the transcription of a Task Force Command Net exercise as the main source of training and test data. The data contain 1400 transcribed utterances which we have divided into two training sets of approximately 500 sentences each and two test sets of approximately 200 sentences each. For the experiments reported here, we make use of only one of the training sets and only one of the test sets. In addition, we had generated 33 sentences within the domain as an initial data set, giving us a total of 530 training sentences.

The bigram and trigram language models were trained from these 530 sentences using standard techniques. TINA's rules were developed by hand, based on observed patterns in these sentences. TINA's probabilities were trained automatically by parsing each training sentence and updating appropriate counts. It should be noted that TINA can only parse and understand 321 of the 530 training sentences (60.6%).[7] The only knowledge TINA has of the other 209 sentences is of the existence of the individual words in these sentences. There are 171 words which appear in those 209 sentences that do not appear in the rest of the training data. Hence, the TINA language model and, by inference, the TINA-LM system and the TINA-generated bigram have no meaningful training data for 171 of CCLINC's 692 words.

### 3.2 Evaluation

We have run very preliminary experiments to obtain initial benchmarks on the performance of the system and its components. In particular, we will report separate results on speech recognition, text understanding, and speech understanding. In all cases, we will be using as the test data one of the unseen sets mentioned above, a set of 190 sentences. For speech recognition, we report for three separate experimental conditions (i.e., distinct language models): data-driven bigram,

[7] We have not yet implemented a robust parsing capability, which would greatly extend TINA's coverage.

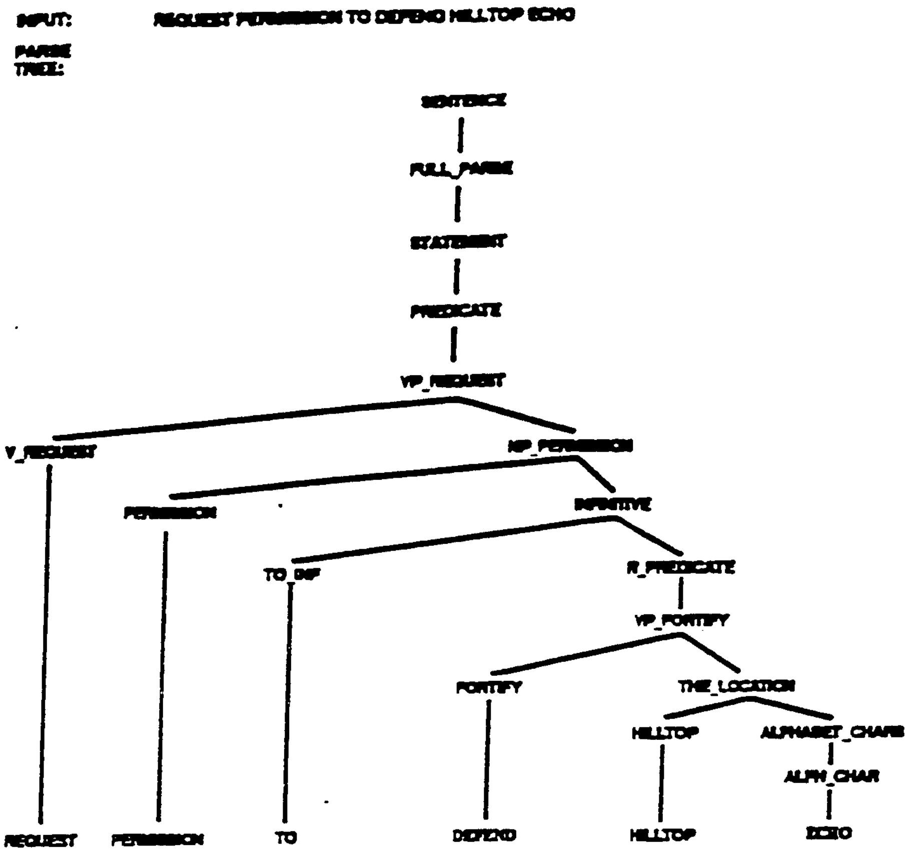


Figure 3: Parse tree for a sample sentence.

```
Input: Request permission to defend hilltop echo

Semantic Frame (Common Coalition Language):
{c statement
   :mode "fpl"
   :number "fpl"
   :pred {p v_request
            :topic {q permission
                      :complement {p fortify
                                     :aux "to"
                            :topic {q hilltop
                               :pred {p initials
                                        :topic "echo" }}}}}}

English Paraphrase: We request permission to fortify hilltop echo
French Paraphrase:  Nous demandons la permission de fortifier le sommet echo
Korean Paraphrase: 우리는 에코고지를 축성하기를 허가를 요구 한다
```

Figure 4: The semantic frame and paraphrases for a sample sentence.

data-driven trigram, and TINA-LM.[8] The performance is evaluated based on insertion, deletion, and substitution error rates as well as word and sentence error rates.

For speech understanding, we also report on the same three conditions. In this case, it is more difficult to measure performance. We decided to adopt the evaluation methodology proposed by White and O'Connell (i.e., fluency and adequacy criteria)[12]. The fluency and adequacy of the French and Korean translations were evaluated by native speakers of those languages. Text understanding was evaluated in the same way except that, in this case, we had only one system.

Table 1 shows the speech recognition results as a function of language model. Note that the sentence error rates are approximately 50% for each of the recognizers. These error rates are higher than expected. We would expect lower error rates if we had used task-specific acoustic models and/or had more training data. As expected, the sentence error rate for the data-driven trigram recognizer is slightly lower than the sentence error rate for the data-driven bigram recognizer. However, the sentence error rate for the TINA-LM recognizer is higher than that of either of the data-driven n-gram recognizers. TINA-LM gives a very high deletion error rate which is due in large part to the near 100% deletion incurred for failed sentences. We show later in this section that, despite higher speech recognition sentence error rates, the TINA-LM system produces "better" translations than do either of the other speech-to-speech translation systems.

The text and speech understanding results are shown in Table 2. The second column of Table 2 indicates the number of test sentences that each system parses (i.e., the number of test sentences for which the system in question produces a parse tree, semantic frame, and paraphrases). The remaining columns of the table show the fluency and adequacy scores of the French and Korean translations, where 1 is the lowest score and 5 is the highest score. The first point to note is that the text translation system parses 52.1% of the 190 test sentences. This is a particularly good result, considering that TINA only parses 57.9% (288/497) of the training sentences taken from the military exercise transcription. The conclusion is that we have covered part of the coalition brigade domain quite well. The second point to note is that the text translation system outperforms the two data-driven n-gram systems, both in terms of number of sentences parsed and number of fluent and adequate parses. This result is, of course, expected since the data-driven n-gram recognizers have high error rates. Another point to note is that the data-driven trigram system does slightly better than the data-driven bigram system. This is also an expected result. Table 2 also shows that the TINA-LM system definitely outperforms the two data-driven n-gram systems. (Note the number of fluent and, in particular, adequate parses for the three systems in question.) In addition, the TINA-LM system performs nearly as well as the text translation system. The TINA-LM French system produces ten fewer adequate parses than does its text translation counterpart and the TINA-LM Korean system produces only one fewer adequate parse than its text translation counterpart. Furthermore, the TINA-LM system parses many more sentences (146 to 99) than does the text translation system. We will discuss this result as well as the general performance and merits of the TINA-LM system in the next section.

There are a number of important caveats to the above experiments. The first and most important caveat is that CCLINC, and therefore any evaluation of it, is still in a preliminary stage. The second caveat is that, as previously mentioned, we only ran a 1-best CSR in our decoupled systems. We would expect the performance of the n-gram systems to improve with the use of N-best CSRs. Finally, TINA's parse coverage on both the training and test sets would improve substantially if we added a robust parsing capability, although the paraphrase quality would probably degrade for robust analyses.

## 4. Discussion

In this section, we shall discuss the merits of the tightly-coupled approach, the portability of CCLINC to new languages, and the applicability of speech translation technology to the coalition brigade domain.

We believe that the TINA-LM system has numerous strengths. First, the system directly incorporates a natural language model into the primary search process of the recognizer. NL constraints are applied immediately in a left-to-right pass through the sentence, thereby coercing the system to produce only grammatical recognizer outputs.[10] Thus, TINA-LM often produces a parseable recognition output even when the output is not correct (i.e., when there is at least one word error in the recognition output). Specifically, the TINA-LM system produces incorrect but parseable recognition outputs for 62 of the 190 test sentences. In contrast, the data-driven bigram system produces incorrect but parseable recognition outputs for only four of the test sentences. It is these numbers which explain how the TINA-LM system produces "better" translations than do the n-gram systems despite higher recognition error rates. These numbers also explain how the TINA-LM system parses more sentences than does the text translation system. In particular, the TINA-LM recognizer transforms 50 unparseable sentences into parseable sentences. In other words, of the 62 test sentences for which the TINA-LM recognizer produces an incorrect but parseable output, only twelve can be parsed by the text translation system. The second strength of the TINA-LM system is that it enforces long-distance language constraints that n-gram language model-based systems can not. For instance,the TINA-LM system correctly recognizes the sentence "Roger I got it." In contrast, the data-driven bigram system produces "Roger I got a" for the same sentence. The output "Roger I got a" does not satisfy the following long-distance, ordering constraint: "... subject verb object end_of_sentence." The third advantage of the TINA-LM system is that it uses a meaning-based generalization mechanism rather than the experience-based generalization mechanism that n-gram language models use. Meaning-based generalization is particularly important when data are sparse, as in our current situation.

One advantage of interlingual systems such as CCLINC is that they are, at least in theory, readily portable to new languages. In practice, we found this statement to be reasonably true. The use of a CCL made extension to French significantly more straightforward since English and French share numerous characteristics. An example of a feature which we needed to add to the CCL to extend CCLINC to French is the ability to distinguish between direct and indirect objects and direct and indirect object pronouns. In English, both objects and object pronouns follow the verb whereas

[8] The TINA-generated bigram was not evaluated because we are not confident that it is bug-free.

[9] A fluent parse is a sentence which is parsed by the appropriate system and whose system translation is given a fluency score of at least three. An adequate parse is defined analogously.

[10] Theoretically, the TINA-LM recognizer should produce a grammatical output for each sentence. However, it may produce no output if there is no sentence hypothesis with the minimum acoustic/linguistic score. In fact, the TINA-LM system did not produce parses for 44 of the 190 test sentences. (See Table 2.)

| Language Model | Substitution Error Rate | Insertion Error Rate | Deletion Error Rate | Word Error Rate | Sentence Error Rate |
|---|---|---|---|---|---|
| Data-driven Bigram | 23.5% | 5.6% | 4.0% | 33.0% | 51.6% |
| Data-driven Trigram | 23.0% | 5.3% | 4.9% | 33.3% | 47.9% |
| TINA-LM | 27.1% | 2.3% | 39.6% | 69.0% | 54.7% |

Table 1: Speech Recognition as a Function of Language Model

| System | Sentences Parsed | Language | Fluency Scores | | | | | Adequacy Scores | | | | | Fluent Parses[9] | Adequate Parses[9] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | | |
| Text Translation | 99 (52.1%) | French | 1 | 0 | 3 | 0 | 95 | 1 | 0 | 4 | 0 | 94 | 98 | 98 |
| | | Korean | 1 | 0 | 1 | 1 | 99 | 2 | 0 | 5 | 2 | 90 | 98 | 97 |
| Data-Driven Bigram | 87 (45.8%) | French | 1 | 0 | 2 | 0 | 84 | 4 | 0 | 1 | 0 | 82 | 86 | 83 |
| | | Korean | 2 | 0 | 0 | 2 | 83 | 2 | 3 | 1 | 1 | 80 | 85 | 82 |
| Data-Driven Trigram | 89 (46.8%) | French | 1 | 0 | 3 | 0 | 85 | 4 | 0 | 2 | 0 | 83 | 88 | 85 |
| | | Korean | 1 | 1 | 0 | 1 | 86 | 3 | 2 | 2 | 1 | 81 | 87 | 84 |
| TINA-LM | 146 (76.8%) | French | 34 | 0 | 9 | 0 | 103 | 58 | 0 | 5 | 0 | 83 | 112 | 88 |
| | | Korean | 13 | 7 | 12 | 6 | 108 | 42 | 8 | 10 | 3 | 83 | 126 | 96 |

Table 2: Text and Speech Translation Results

in French, direct and indirect objects follow the verb, but direct and indirect object pronouns precede the verb.

The use of a CCL made extension to Korean somewhat easier. We did not need to capture "rank" information in the CCL because CCLINC assumes one mode of speaking. (One big difference between English and Korean is that Korean has different verb endings depending on the ranks of the speaker and the listener. CCLINC emulates the speech that an educated military person of middle rank would use to his peers[14].)

Finally, we would like to comment on the applicability of speech translation technology to the coalition brigade domain. In other words, we are interested in how easy it is to automatically translate "military" sentences as compared to sentences in other domains. On the one hand, as much as 40% of our data involves nothing more than user or grid identification or other basic Army protocols. On the other hand, "militarese" is more ungrammatical and colloquial than is typical speech. Furthermore, it is difficult to find translators and evaluators with military knowledge, both of which are needed in the development of CCLINC.

## 5. Future Plans

Based on our initial results and an assessment of user needs in Korea, we expect that the focus of our work in the near future will be on language modeling and understanding of real message traffic, which will serve as a basis for application to both text and speech translation.

## References

[1] J. Glass, D. Goodine, D. Phillips, M. Sakai, S. Seneff, and V. Zue, "A Bilingual VOYAGER System," Proc. Eurospeech, Berlin, Germany, 1993.

[2] J. Glass, J. Polifroni, and S. Seneff, "Multilingual Language Generation Across Multiple Domains," Proc. ICSLP, Tokyo, Japan, 1994.

[3] E. Hovy, Machine Translation. Session Summary, Proc. Human Language Technology Workshop, Plainsboro, NJ, March 1994.

[4] F. Kubala, J. Bellegranda, J. Cohen, D. Pallett, D. Paul, M. Phillips, R. Rajasekaran, F. Richardson, M. Riley, R. Rosenfeld, B. Roth, and M. Weintraub, "The Hub abd Spoke Paradigm for CSR Evaluation," Proc. Human Language Technology Workshop, Plainsboro, NJ, March 1994.

[5] D. B. Paul, "A CSR-NL Interface Architecture," Proc. ICSLP 92, Banff, Alberta, Canada, Sept. 1992.

[6] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," Proc. ICSLP 92, Banff, Alberta, Canada, Sept. 1992.

[7] D. B. Paul and B. F. Necioglu, "The Lincoln Large-Vocabulary Stack-Decoder HMM CSR," ICASSP 93, Minneapolis, April 1993.

[8] D. B. Paul, "New Developments in the Lincoln Stack-Decoder Based Large-Vocabulary CSR System, to be submitted to the ARPA Spoken Language Technology Workshop, Austin, Texas, Jan. 1995.

[9] M. Rayner, H. Alshawi, I. Bretan, D. Carter, V. Digalakis, B. Gamback, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price, and C. Samuelsson, "A Speech to Speech Translation System Built From Standard Components," Proc. Human Language Technology Workshop, Princeton, NJ, March 1993.

[10] D. Roe, F. Pereira, R. Sproat, and M. Riley, " Toward a Spoken Language Translator for Restricted-Domain Context-Free Languages," Proc. Eurospeech, Berlin, Germany, 1991.

[11] S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, vol. 18, no. 1, 1992.

[12] J. White and T. O'Connell, "Evaluation in the ARPA Machine Translation Program: 1993 Methodology," Proc. Human Language Technology Workshop, Plainsboro, NJ, March 1994.

[13] M.Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, and W. Ward, "Recent Advances in JANUS: A Speech Translation System," Proc. Human Language Technology Workshop, Princeton, NJ, March 1993.

[14] D. Yang, personal communication.

[15] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill, "PEGASUS: A Spoken Language Interface for On-Line Air Travel Planning," Proc. Human Language Technology Workshop, Plainsboro, NJ, March 1994.

## Forward Area Language Converter

Mr. Daniel W. Smith, Jr.

- Initial prototype system will demonstrate translation of 2-3 languages.

- Final System will include language translation capabilities to support XVIII Airborne Corps contingencies.

- System user-friendly utilizing a Graphical User Interface (GUI).

- Final version of system software will step the soldier through the document scanning procedure. Once document is scanned, the soldier will essentially "press a key" and initiate an automatic OCR/translation procedure of the scanned information followed by transmission over a SINCGARS radio or the MSE digital communications systems. Custom integration software will take care of all the necessary calls to the program, file generation, execution, etc., this procedure will be transparent to the user.

Contact: Mr. Daniel W. Smith, Jr.
Science Advisor
CDR XVIII Airborne Corps
ATTN: AFZA-CS-S
Ft. Bragg, NC 28307-5000
(910) 396-3780; FAX: (910) 396-8215

# FALCON

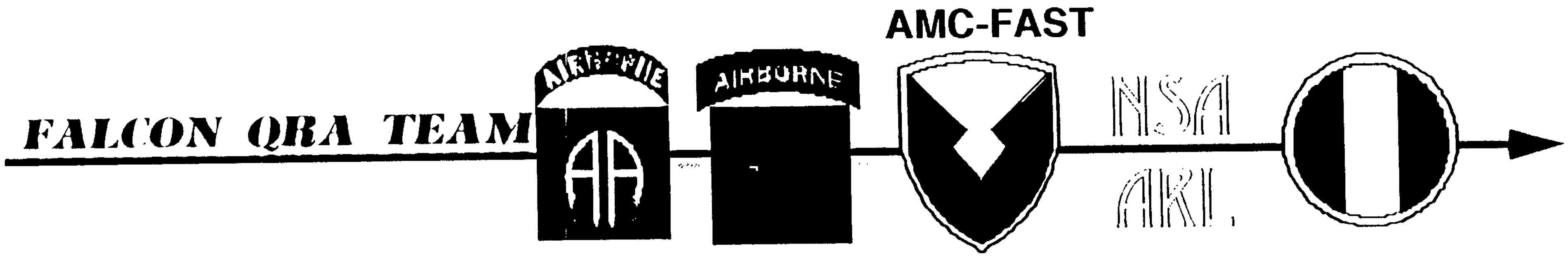
## Forward Area Language Converter

prepared by:
Perri Nejib, Asst. Science
Adviser, XVIII Airborne Corps,
910-396-3780

# Forward Area Language Converter

**WARFIGHTER SHORTFALL:**

The 82nd Airborne Division has no portable, tactical capability for performing a language translation on documents in foreign languages. This operational deficiency was apparent in Haiti during OPERATION UPHOLD DEMOCRACY, where trained linguists were scarce or assigned to higher priority missions. This is a recurring problem noted in JUST CAUSE and OPERATION DESERT SHIELD/STORM reports.



# Forward Area Language Converter

**QRP STATUS:**

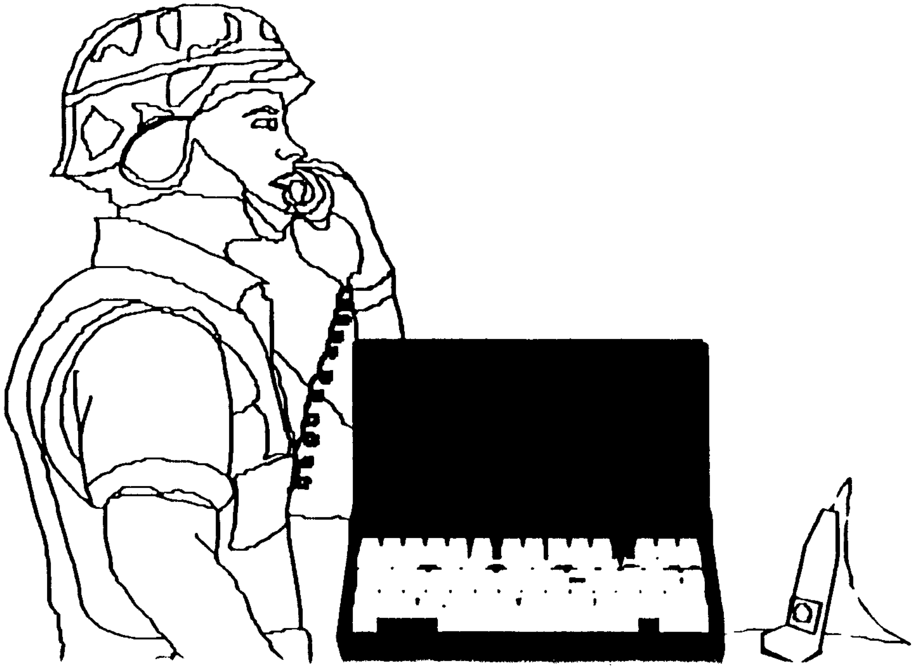
- Operational Needs Statement (ONS) approved by Division Commander, 82d Airborne Division, (19 May 95)

- Funding acquired from AMC-FAST office and TRADOC EELS Battle Lab

- 82nd Airborne Proponent G2

- Proof of Concept Demo scheduled for June 95.

Testing scheduled in Haiti and Panama, August/Sept 95



# Forward Area Language Converter

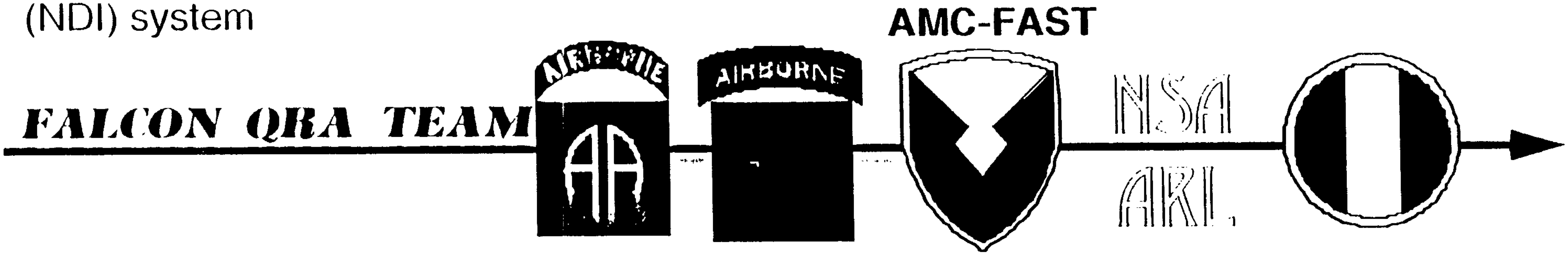
**SOLUTION:**

- Provide an Operational Need Statement (ONS)

- FALCON (Machine Translation) Working group formed 02 May 95 consisting of 82nd Airborne G2, XVIII Airborne Corps Science Adviser (AMC-FAST), NSA, Army Research Lab, and TRADOC EELS Battle Lab
  - Future working group meetings will also include Rome Labs

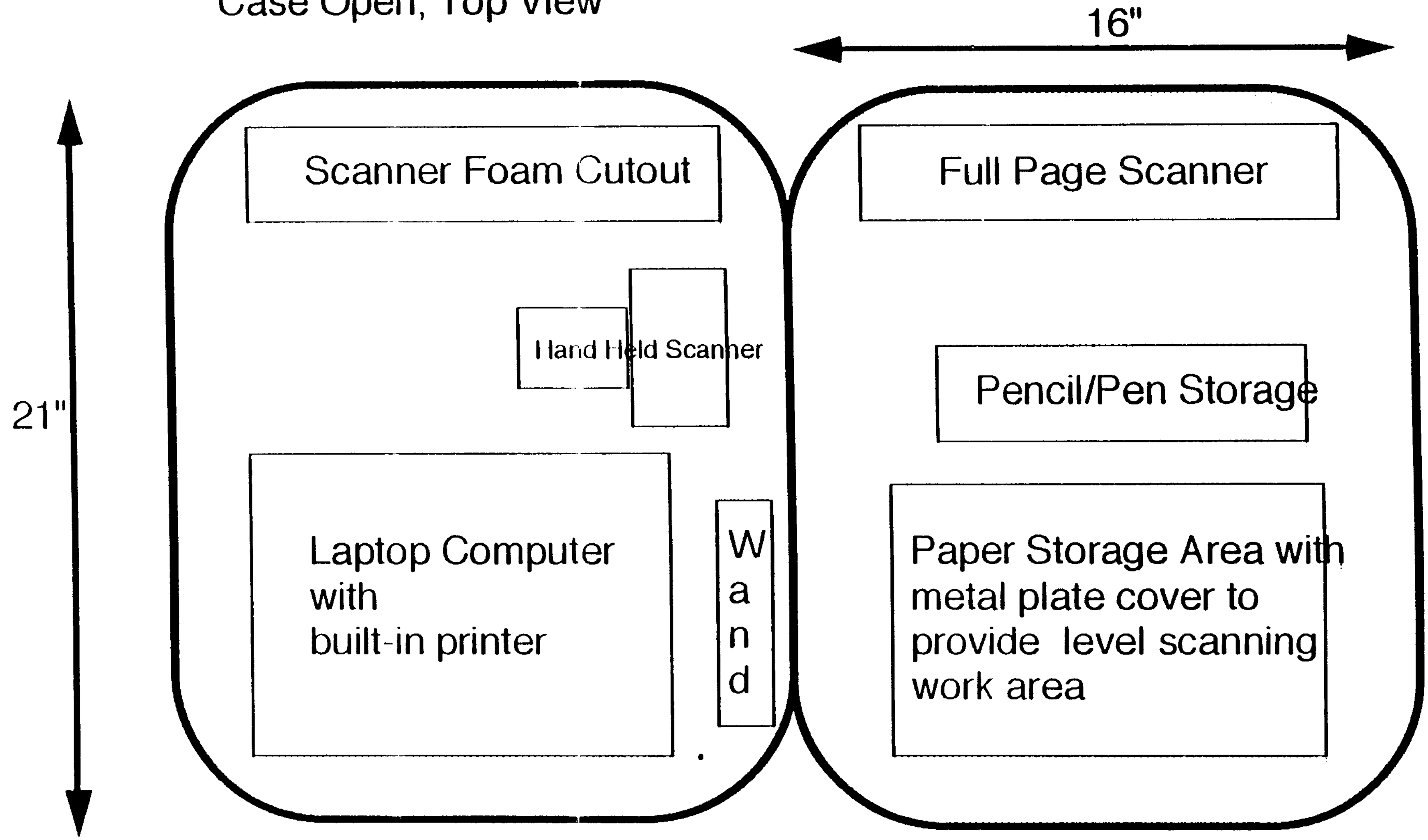
- Leverage government developed technology in the areas of Optical Character Recognition (OCR), Machine Language Translation and digital communications systems to develop a portable Language Translator for soldier and marine use

- On a quick reaction procurement rapidly develop a non-developmental item (NDI) system



# FALCON System Layout

Case Open, Top View



16"

21"

Scanner Foam Cutout

Hand Held Scanner

Laptop Computer
with
built-in printer

W
a
n
d

Full Page Scanner

Pencil/Pen Storage

Paper Storage Area with
metal plate cover to
provide level scanning
work area

# FALCON System Specifications and Description

- Initial prototype system will demonstrate translation of 2-3 languages

- Final System will include language translation capabilities to support XVIII Airborne Corps contingencies

- System user-friendly utilizing a Graphical User Interface (GUI).

- Final version of system software will step the soldier through the document scanning procedure. Once document is scanned, the soldier will essentially "press a key" and initiate an automatic OCR/translation procedure of the scanned information followed by transmission over a SINCGARS radio or the MSE digital communications system. Custom integration software will take care of all the necessary calls to the program, file generation, execution, etc., this procedure will be transparent to the user.

# FALCON SYSTEM

- Candidate for joint service use and dual technology by U.S. Government and commercial agencies and industry

# Operational Concept

**Short Term Solution: (June 95)**

- Will allow a soldier working at a checkpoint to translate a captured or otherwise received foreign language document into English and print the complete document on a printer followed by voice comunications over a tactical radio to higher headquarters.

**Follow-On Effort: (Dec/Jan 96 - 2nd Gen System)**

- Focus on miniaturization and the introduction of new technology as it becomes available in this high priority R&D area e.g. TCIM (Tactical Communications Interface Module)).

**Long Term Goal:**

- Downsize system to a small hand-held device that will fit into a soldier's BDU cargo pocket.

AMC-FAST

***FALCON QRA TEAM***

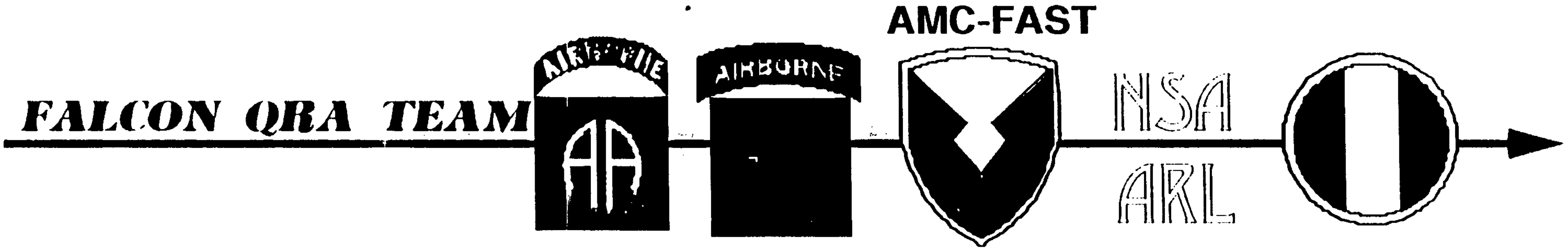
AIRBORNE

NSA

ARL

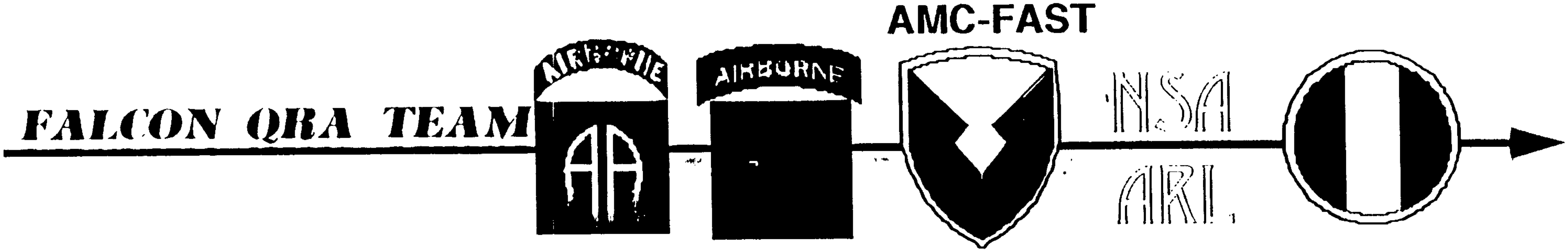
# Organizational Concept

- A prototype FALCON system will be provided to G2, 82nd Airborne Division for field test and evaluation
- FALCON systems would be issued to combat battalions with additional systems provided to Military Police and Military Intelligence units.
- Control of the FALCON system would be in accordance with control procedures for existing military intelligence assets



*QRA OF THE FALCON SYSTEM CLEARLY DEMONSTRATES ...*

FEDERAL LABORATORY SYSTEM MEETING THE NEEDS OF WARFIGHTERS!!!



## Multimedia Medical Language Translator

HMC(AW) Michael D. Hesslink
Captain Michael Valdez

The Multimedia Medical Language Translator (MLT) uses a laptop computer to help medial examiner communicate with patients. The system enables a health-care provider to ask a series of standard examination questions, and to convey simple words of greeting and explanation, in a patient's native tongue. This contact can make all the difference in keeping the patient calm and in getting the information necessary to prompt, effective treatment.

Developed by Commander Lee Morin of the U.S. Navy Medical Corps, MLT was first used by U.S. Navy health=care staff of Fleet Hospital Zagreb, while supporting U.N. peacekeeping forces in the former Yugoslavia. The hospital is responsible for the health care of 40,000 U.N. personnel from 35 nations.

Distributed as a CD-ROM disk, the program is applicable to any type of health-care environment. It promises to be especially valuable i crises--such as natural disasters or political conflicts, or in emergency rooms of metropolitan hospitals -- where rapid response is needed and interpreters may not be readily available.

The current version of MLT can be used by anyone literate in English, Russian, or Chinese. He or she can point to a series of phrases from a list of nearly 2,000 or select one of more than 40 "scripts" for various topics and specialties, from dentistry to gynecology. The device then "speaks" the phrases or script in the voice of a native speaker form one of several dozen languages. One script cycles through all available languages, asking the patient, "Do you speak...?" The medical worker can also use the computer's search function to instantly find desired words or phrases.

Written in state-of-the-art Visual Basic running under Microsoft Windows, the MLT program is compact and can function on a basic machine with 4 megabytes of RAM and a single-speed CD player. The device can be customized to each user.

Contact:
HMC(AW) Michael D. Hesslink
Naval Aerospace and Operational Medical Institute
ATTN: Code 05
220 Hovey Rd.
Pensacola, FL 32508-1047
(904) 452-8212; FAX: (904) 452-3404

Technology Review:
Special Operations Forces (SOF)
Speech Recognition for Language Sustainment

# Appendix D:
# References

# References

Levin, L., Glickman, O., Qu, Y., Gates, D., Lavie, A., Rose, C., Ess-Dykema, C., & Waibel, A. (1995). Using Context in Machine Translation of Spoken Language. Carnegie Mellon University and U.S. Department of Defense. In Proceedings of the Theoretical and Methodical Issues in Machine Translation Conference. Leuven, Belgium, July 5-7, 1995.

Montgomery, C.A., Stalls, B.G., Belvin, R.S., Arnaiz, A.R., Stumberger, R.E., Li, N., Litenatsky, S.H. (1995). Machine-Aided Voice Translation (MAVT): Advanced Development Model. Reprinted from the 5th Annual IEEE Dual-Use Technologies and Applications Conference, May 22-25, 1995, p. 112-118.

Nejib, P. (1995). Forward Area Language Converter. XVIII Airborne Corps, Ft. Bragg, NC.

Spoken Language Systems Group (1995). Research and Development of Multilingual Conversational Systems. Laboratory for Computer Science, Massachusetts Institute of Technology, August 2, 1995.

Suhm, B. Geutner, P., Kemp, T., Lavie, A., Mayfield, L., McNair, A., Rogina, I., Schultz, T., Sloboda, T., Ward, W., Woszczyna, M., and Waibel, A. (1995). JANUS: Towards Multilingual Spoken Language Translation. Carnegie Mellon University (USA), and Karsruhe University (Germany). In Proceedings of the ARPA Spoken Language Technology Workshop. Austin, TX, January 1995.

Tummala, D., Seneff, S. Paul, D., Weinstein, C., and Yang, D. (1995). CCLINC: System Architecture and Concept Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications. Lexington, MA: Lincoln Laboratory, MIT. In Proceeding of the ARPA Spoken Language Technology Workshop.

# Technology Review:
# Special Operations Forces (SOF)
# Speech Recognition for Language Sustainment

## Appendix E:
## Revised List of Participants

## Revised List of Participants

Mr. Ray Lane Aldrich
HQ Dept. of the Army
ODCSINT - Pentagon, Rm. 2B479
Washington, DC 20310-1001
(703) 695-2120; FAX: (703) 693-2038
e-mail: aldrichl@pentagon-hqdadss.Army.mil

Eladia Arroyo
3rd Special Forces Group (Airbome)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

Dr. Madeleine Bates
BBN Systems & Technologies
70 Fawcett St.
Cambridge, MA 02128
(617) 873-3634; FAX: (617) 873-2534
e-mail: bates@bbn.com

Dr. Jared Bemstein
Entropic Research Laboratory
1040 Noel Drive
Menlo Park, CA 94025
(415) 328-8877; FAX: (415) 328-8866
email: jared@entropic.com

Mr. Brian Berrey
WinTee, Inc.
12805 Old Fort Rd.
Ft. Washington, MD 20744
(301) 203-0774; FAX: (301) 203-8049
bberrey@ostgate.com

Dr. Deniz I. Bilgin
Defense Language Institute
Foreign Language Institute
ATTN: ATFL-DCI-TI, Bldg 635
Presidio of Monterey, CA 93944-5006
FAX: (408) 242-6466

Dr. Frank L. Borchardt
Duke University
Department of German
116H Old Chem, Box 90256
Durham, NC 27708-0256
(919) 660-3161; FAX (919) 660-3166
e-mail: frankbo@acpub.duke.edu

LTC Robert Brady
US Army Special Forces Command, G-3
Fort Bragg, NC 28307-5000
(910) 432-7511

Dr. Barbara D. Broome
U.S. Army Research Laboratory
ATTN: AMSRL-IS-TP
Aberdeen Proving Ground, MD 21005-5067
(410) 278-4773/4196; FAX: (410) 278-4204
e-mail: bdbroome@arl.mil

Mr. Gilbert W. Buhrmann, Jr.
Senior Project Engineer
United States Special Operations Command
Office of Special Technology
10530 Riverview, Building 3
Fort Washington, MD 20744-5821
(301) 203-2670; FAX: (301) 203-2641

LTC Carlos A. Burgos
HHC, 7th Special Forces Group (A)
Ft. Bragg, NC 28308-5000
(910) 432-1809

Dr. Bill Byme
CLSP, Johns Hopkins University
3100 N. Charles St.
Baltimore, MD
(410) 516-4120
e-mail: byme@jhu.edu

Dr. Beth Carlson
MIT Lincoln Laboratory
244 Wood Street, Rm S4-115
Lexington, MA 02173-9108
(617) 981-5375; FAX: (617) 981-0186
e-mail: BETH@SST.LL.MIT.EDU

Mr. Paul R. Chatelier
OSTP-CAETI (ARPA)
1901 Beauregard St., Suite 510
Alexandria, VA 22331
(703) 998-1313; FAX: (703) 379-3778
e-mail: pchat@dmso.dtic.dla.mil

Mr. George Chen
SRI International
33 Ravenswood Avenue
Menlo Park, CA 94025
(415) 859-2204; FAX: (415) 859-5984
e-mail: gtchen@speech.sri.com

Dr. Ray Clifford, Provost
Defense Language Institute
Defense Language Center
Presidio of Monterey, CA 93944-5006

LTC James H. Coffman, Jr.
HQDA, ODCSOPS, DAMO-ODP
400 Army Pentagon
Washington, DC 20310-0400
(703) 697-3578; FAX (703) 614-5014

Mr. Sean Colbath
BBN Systems & Technologies
70 Fawcett St.
Cambridge, MA 02128
(617) 873-3847; FAX: (617) 873-2534
e-mail: scolbath@bbn.com

Dr. Raymond Cook, ORD
U.S. Department of the Army
131 Governors Drive
Leesburg, VA 22075
(202) 965-3517; FAX: (703) 243-4127

Ms. Molly Cruz
USA JFK Special Warfare Center & Sch
Language Development Center
3d Bn, 1st SWTG (A)
Bldg D-3206, Room 305
Fort Bragg, NC 28307-5000
(910) 432-4400; FAX: (910) 432-6511
e-mail: 3bn-S32@usasoc.soc.mil

Mr. Russ Dube
Interactive Drama Inc.
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 10814
(301) 654-0676; FAX: (301) 657-9174

Dr. Kathleen Egan
Office of Research and Development/
Interagency Technology Office
1250 Maryland Avenue, SW, #6300
Washington, DC 20202-5544
(202) 708-5542/6001; FAX (202) 708-6

Dr. Maxine Eskenazi
Carnegie Mellon University
215 Cyert Hall, Robotics Institute
Pittsburgh, PA 15213-3890

Ms. Lisa C. Frey
Camber Corp.
601 13th St., NW - Ste. 350 North
Washington, DC 20005
(202) 393-1648; FAX: (202) 628-8498

Mr. Bernard Greene
HumRRO
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314
(703) 549-3611; FAX: (703) 549-9025

Dr. John Gurney
Army Research Laboratory
3505 Kensington Court
Kensington, MD 20895
(301) 394-3920; FAX: (301) 394-3903
email: gurney@adelphi-assbo1.arl.mil

Ms. Nadine A. Hadge
Digital Systems Research, Inc.
4301 N. Fairfax Drive, Suite 725
Arlington, VA 22203
(703) 522-6067, ext. 157; FAX: (703) 522-6367
WINS%"<nhadge@ssto.snap.org>"

Mr. Martin R. Hall
Johns Hopkins University
Applied Physics Laboratory
Johns Hopkins Road
Laurel, MD 20723-6099
(301) 953-6221; FAX: (301) 953-6904
e-mail: marty hall@jhuapl.edu.

Dr. William G. Harless, President
Interactive Drama Inc.
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 10814
(301) 654-0676; FAX: (301) 657-9174
e-mail: intdrama@aol.com

Kerry Heinricht, ISCS
Naval Special Warfare Development Group
ATTN: CMD Lang Prgm Mgr (N3LP)
1636 Regulus Avenue
Virginia Beach, VA 23461-2299
(804) 433-7960, x177; FAX: (804) 433-7960, x377
(no e-mail)

COL Woody Held
Dept of Foreign Languages
U.S. Military Academy
West Point, NY 10996
(914) 938-5286; FAX: (914) 938-3585

HMC(AW) Michael D. Hesslink
Naval Aerospace & Operational Medical Institute
ATTN: Code 05
220 Hovey Road
Pensacola, FL 32508-1047
(904) 452-8212; FAX: (904) 452-3404

Dr. David Hislop
Army Research Office
Box 12211
Research Triangle Park, NC 27709-2211
e-mail: <HISLOP@aro-emh1.army.mil>

Dr. Melissa Holland
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(703) 274-5569; FAX: (703) 274-3573
e-mail: holland@alexandria-emh2.army.mil

Ms. Helena Hughes
Federal Language Training Laboratory
801 N. Randolph St., Suite 201
Arlington, VA 22203
(703) 525-4287; FAX: (703) 525-5186

Mr. Fred Jacome
Duke University
Humanities Computing Facility
Box 90269, 015 Language Building
Durham, NC 27708-0269
(919) 660-3192; FAX: (919) 660-3191
e-mail: jacome@acpub.duke.edu

Suzette M. Jadotte
3rd Special Forces Group (Airborne)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

SFC Michael P. Judge
US Army Intelligence Center
Headquarters, USAIC & FH
ATTN: ATZS-TPS-L
Ft. Huachuca, Arizona 85613-6000
(520) 533-2360; FAX: (520) 538-8744

Dr. Jonathan Kaplan
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(703) 274-8828; FAX: (703) 274-3575
e-mail: kaplan@alexandria-emh2.army.mil

LTC Victor Kjoss
DCSOPS, Training Division
Fort Bragg, NC 28307-5000
(910) 432-8720

Dr. Elizabeth Klipple
University of Maryland
Department of Computer Science
4800 Berwyn House Rd., #104
College Park, MD 10740
(301) 405-2716; FAX: (301) 405-6707
e-mail: Klipple@cs.umd.edu

Dr. Mazie Knerr
HumRRO
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314
(703) 706-5634; FAX: (703) 549-9025
e-mail: knerrm@alexandria-emh2.army.mil

Dr. Gregory M. Kreiger
Deputy Assistant Commandant
U.S. Army Intelligence Center
ATTN: ATZS-DAC
Ft. Huachuca, AZ 85613-6000
(520) 538-7303; FAX: (520) 538-7409
e-mail: kreigerg@huachuca-emh11.army.mil

Dr. Anita Kulman
10107 Snowden Road
Laurel, MD 20708
(301) 688-8901

LTC Steve LaRocca
Department of Foreign Languages
U.S. Military Academy
West Point, NY 10996
(914) 938-5286; (FAX) (914) 938-3585
e-mail: gs0416@usma3.usma.edu

Dr. Young-Suk Lee
MIT Lincoln Laboratory
244 Wood St., Rm. S4-113
Lexington, MA 02173-9108
(617) 981-2703; FAX: (617) 981-0186
e-mail: ysl@sst.LL.mit.edu

Mr. Chris Lindstrom
18th ABC G-2
703 Larkspur Drive
Fayetteville, NC 28311
(910) 396-4126/5803; FAX: (910) 396-3

Dr. Susann Luperfoy
MITRE Corporation
7525 Colshire Drive
McLean, VA 22102
(703) 883-6091; FAX: (703) 883-6435
e-mail: susann@azrael.mitre.org

Dr. Jack Lynch
MIT Lincoln Laboratory
244 Wood St. - Rm S4-177
Lexington, MA 02173-9108
(617) 981-2746; FAX: (617) 981-0186
e-mail: JTL@SST.LL.MIT.EDU

Dr. Arthur McNair
Carnegie Mellon University
School of Computer Science
5000 Forbes Avenue
Pittsburgh, PA 15213
(412) 268-1411; FAX: (412) 268-5578
e-mail: arthurem@cs.cmu.edu

Mr. Louis Meza
7th Special Forces Group (Airborne)
Language Training Facility
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-8857

Mr. Michael Miller
USSOCOM
Commander in Chief
HQ, U.S. Special Operations Command
ATTN: SOSD-SA (Mr. Miller)
7701 Tampa Point Blvd.
MacDill Air Force Base, FL 33621-5323
(813) 840-5285; FAX: (813) 840-5266

Dr. Christine A. Montgomery
Language Systems, Inc. (LSI)
6269 Variel Ave., Suite F
Woodland Hills, CA 91367
(818) 703-5034; FAX: (818) 703-5902
e-mail: chris@lsi.com

Dr. Jack Mostow
Director, Project LISTEN
Carnegie Mellon University
215 Cyert Hall, Robotics Institute
Pittsburgh, PA 15213-3890
(412) 268-1330; FAX: (412) 268-6298
e-mail: "mostow@cs.cmu.edu"

Mr. Leo Neumeyer
SRI International
33 Ravenswood Avenue
Menlo Park, CA 94025
(415) 859-4522; FAX: (415) 859-5984
e-mail: leo@speech.sri.com

CPT Edward Nickerson
HHD 525 MI Brigade
Ft. Bragg, NC 28307-5000
(910) 396-9301/5266; FAX (910) 396-4647

Mr. David Nicks
Interactive Drama Inc.
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 10814

Mr. Glen H. Nordin
DCI Foreign Language Committee
Community Management Staff
Washington, DC 20505
(703) 482-2677; FAX: (703) 482-0684

Dr. Dale E. Olsen
The Johns Hopkins University Applied
Physics Laboratory
Johns Hopkins Road
Laurel, MD 20723-6099
(301) 953-6869; FAX: (301) 953-6682

Mr. John Parker
Rome Laboratory (USAF)
RL/IRAA
32 Hangar Road
Griffiss AFB, NY 13441-4114
(315) 330-4025; FAX: (315) 330-2728
e-mail: parkerj@rl.af.mil

LTC Boyd D. Parsons, Jr.
Joint Special Operations Forces Institute
P. O. Box 71929
Ardennes St.
Ft. Bragg, NC 28307-5000
(910) 432-4509; FAX (910) 432-5467

Mr. Joseph S. Pereira
USA JFK Special Warfare Center & School
Language Development Center
3d Bn, 1st SWTG (A)
Bldg D-3206, Room 305
Fort Bragg, NC 28307-5000

Therse X. Pham
3rd Special Forces Group (Airborne)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

Ms. Jacqueline Pogany
Federal Language Training Laboratory
801 N. Randolph St., Suite 201
Arlington, VA 22203
(703) 525-4473; FAX: (703) 525-5186

Dr. Joseph Polifroni
Spoken Language Systems Group
MIT
545 Technology Square
Cambridge, MA 02139
(617) 253-0248; FAX: (617) 258-8642
e-mail: joe@lcs.mit.edu

Dr. Patti Price
SRI International
33 Ravenswood Avenue
Menlo Park, CA 94025
(415) 859-5845; FAX: (415) 859-5984
e-mail: pprice@speech.sri.com

SFC Lester Pruitt
3rd Special Forces Group (Airborne)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

Mr. Sal Raineri
CMDR, USASOC
Attn: DCSRI-SOFCIL (Mr. Raineri)
Ft. Bragg, NC 28307
910-396-5456/7608

Ms. Florence Reeder
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Jorge Rios
George Washington University
Medical School
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 10814

Mr. Calvin B. Rome
SOF Language Office, DCSOPS, USAS(
854 Danish Dr.
Fayetteville, NC 28303
e-mail: dtd-LO3@soc.mil

Dr. Jim Rorke
US Army Special Forces Command, G-3
ATTN: AOSO-GCT-I
Ft. Bragg, NC 28307-5000
(910) 432-6980; FAX: (910) 432-8050
e-mail: HSKK86A@prodigy.com

SFC Thomas L. Rosenbarger
U.S. Army 5th Special Forces Group (A
ATTN: GRP S-5, ACOIC
Fort Campbell, Ky 42223-5000
(502) 798-7713

Dr. Martin Rothenberg
Syracuse Language Systems
719 E. Genesee St.
Syracuse, NY 13210
(315) 478-6729/800 688-1937; FAX: (3
6902

SGT James A. Rudolf
1st Special Forces Group (Airborne)
ATTN: AOSO-SFI-SC
Fort Lewis, WA 98433-7000
(206) 967-8639; FAX: (206) 357-8669

Dr. Marikka Rypa
SRI International
33 Ravenswood Avenue
Menlo Park, CA 94025
(415) 859-3648; FAX: (415) 859-5984
e-mail: marikka@speech.sri.com

Grumm Victoria Saenz
3rd Special Forces Group (Airborne)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

Dr. Michael G. Sanders
ATTN: Psychology Section
P.O. Box 70660
US Army Research Institute
Fort Bragg, NC 28307-5000
(910) 396-0874; FAX: (910) 396-1102

Mr. J. Allen Sears
Advanced Research Projects Agency/SISTO
3701 North Fairfax Drive
Arlington, VA 22203-1714
(703) 696-2259; FAX: (703) 696-0564
e-mail: asears@arpa.mil

Dr. Robert J. Seidel
Chief, Advanced Training Methods
U.S. Army Research Institute
ATTN: PERI-II
5001 Eisenhower Avenue
Alexandria, VA 22333
(703) 274-8838; (703) 274-3575
e-mail: seidel@alexandria-emh2.army.mil

Dr. Stephanie Seneff
Spoken Language Systems Group
MIT
545 Technology Square
Cambridge, MA 02139
(617) 253-0451; FAX: (617) 258-8642
e-mail: seneff@lcs.mit.edu

Mr. Daniel W. Smith, Jr.
Science Advisor
CDR XVIII Airborne Corps
ATTN: AFZA-CS-S
Ft. Bragg, NC 28307-5000
(910) 396-3780; FAX: (910) 396-8215

Mr. Robert Stumberger
Language Systems Inc.
6269 Varial Ave., Suite F
Woodland Hills, CA 91367
(818) 703-5034/818; FAX: (818) 703-5902

CSM William P. Traeger
Senior Enlisted Advisor
United States Special Operations Command
Joint Special Operations Forces Institute
Bldg D-2507
Fort Bragg, NC 28307-5000
(901) 432-1727; FAX: (901) 432-5467

Mr. Michael Valatka
Office of Research and Development
Mail Stop 4122
Washington, DC 20505
(703) 351-2763; FAX: (703) 243-4127

Captain Michael Valdez
Naval Aerospace & Operational Medical Institute
220 Hovey Road
Pensacola, FL 32508-1047
(904) 452-8212; FAX: (904) 452-3404

Dr. Alex Waibel
Carnegie Mellon University
School of Computer Science
5000 Forbes Ave.
Pittsburgh, PA 15213
(412) 268-7676; FAX: (412) 268-5578
e-mail: ahw@cs.cmu.edu

Ms. Sharon M. Walter
Rome Laboratory (USAF)
RL/IRAA2
32 Hangar Road
Griffiss AFB, NY 13441-4114
(315) 330-4025; FAX: (315) 330-2728
e-mail: walter@ai.rl.af.mil

SFC Nick Ward
U.S. Army 5th Special Forces Group (Airborne)
ATTN: C Company, 3/5 SFG (A)
Fort Campbell, KY 42223-6207
(502) 798-6983; FAX: (502) 798-4115

Dr. Clifford Weinstein
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02173-9108
(617) 981-7491; FAX: (617) 981-0186
e-mail: CJW@SST.LL.MIT.EDU

LTC Harold E. Williams
USA JFK Special Warfare Center and School
Fort Bragg, NC 28307-5000
(910) 432-4400

Mr. William T. Williams
USSOCOM/SOSD-T
7701 Tampa Point Blvd.
MacDill AFB, FL 33621-5323
(813) 840-5263; FAX: (813) 840-5266

Mr. Edward Wolcoff
Camber Corporation
5203 Leesburg Pike, Suite 807
Falls Church, VA 22041
(703) 931-1388; FAX: (703) 931-1393
e-mail: wolcoff@otsg-amedd.army.mil

Mr. Gary Wright
U.S. Army, HQ TRADOC, TDAD
HQ TRADOC
ATTG-CF (ATTN: Gary Wright)
Ft. Monroe, VA 23651-5000
(804) 728-5532; FAX: (804) 728-5544

Isis D. Yoseaf
3rd Special Forces Group (Airborne)
Kuwait Drive
Ft. Bragg, NC 28308-5000
(910) 396-6687; FAX: (910) 396-3903

Dr. Torsten Zeppenfeld
Carnegie Mellon University
Dept of Computer Science
5000 Forbes Ave.
Pittsburgh, PA 15213

Dr. Marcia A. Zier
Interactive Drama Inc.
7900 Wisconsin Avenue, Suite 200
Bethesda, MD 10814

Dr. Victor W. Zue
MIT
Spoken Language Systems Group
545 Technology Square
Cambridge, MA 02139
(617) 253-8315; FAX: (617) 258-8642
e-mail: zue@mit.edu